CHAPTER 6

STATISTICAL MODELS TO PREDICT THE PINE CATERPILLAR OUTBREAK

The models were developed by using the relationship between some dependent variables and independent variables those were defined as below. For the multiple linear regression, the dependent variable was ***larvae density in December*** that is herein referred by LARDEN. For the binary logistic regression, the dependent variable was ***Logit*** or ***Log of odds of epidemic occurrence*** that was defined in equation [4.11] and [4.12]. The independent variables are forest age level 1, 2, 3, 4, and 5, which are herein read as FORAGE1, FORAGE2, FORAGE3, FORAGE4, and FORAGE5, respectively.

6.1. The multiple regression of larva density in December

6.1.1 The models developed by considering all weather factors in each month

The models were developed by considering the same weather factors in each month. These models were developed to identify what month had the closest relationship with the density of larvae in December if the same weather factors are involved in the models. The results of these models were shown in Table 6.1. Six models were developed in which the same weather factors were taken into account in each month from June to November.

Table 6.1 shows that the forest age was one of the factors that affected to the larvae density in December. In which, the forest age level 5 (greater than 20 years old) was the reference category. In all the models, the regression coefficient of forest age level one was not significant at level 5%, so it was removed from the models. It means that the larvae density of the forest age level one was the same as level five. For other levels of forest age, it can be seen that the regression coefficients were

positive. It means that the larvae density of these levels of forest were higher than the forest age level five. Moreover, the highest larvae density should be the forest age level three, followed by the forest age level two, and the last one is forest age level four, that were determined by the values of the regression coefficients. The larvae density in forest age level two, three, and four is higher than that of level five at the minimum value is 6.3, 8.7, and 2.0 larvae per tree respectively, and the higher maximum is 7.0, 9.6, and 2.4 larvae per tree for the forest age level two, three and four, respectively.

Weather factors involved in the models were different in different months. The model which was developed in August (model 1.1.3) involved least weather factors with only one weather factor. It was the total of rainfall in August. While other models were involved more weather factors. The models, which were developed in September and October, weather data involved five weather factors. However, all the developed model were significant at the level less than 0.0005 by using the F test these results were shown in the last two columns in Table 6.1. Among seven weather factors, the number of rainfall days factor was removed from all models. It means that this factor was not significant when other factors were used.

The weather factors in different month had varying effects on larvae density as indicated by different values of regression coefficient. The average humidity and the highest temperature had negative affect in this period because the regression coefficients of these factors were negative. It means that when the highest temperature or average humidity increases the larvae density decreased. Other factors can be a negative or positive depending on months.

Table 6.1: Regression models for larvae density in December in relation to forest age and all weather factors in each month.

Model	Equation of regression		R	R^2	SE	F	Sig.
1.1.1 (June)	LARDEN =	45.228 + 6.944 FORAGE2 + 9.622 FORAGE3 + 2.433 FORAGE4 - 0.031 JUN.SUNH - 0.502 JUN.HUMID + 0.012 JUN.RAIN	0.63	0.40	5.34	24.15	0.000
1.1.2 (July)	LARDEN =	33.772 + 6.580 FORAGE2 + 9.273 FORAGE3 + 2.375 FORAGE4 + 0.025 JUL.RAIN - 0.364 JUL.HUMID + 0.035 JUL.SUNH - 0.411 JUL.TMAX	0.66	0.44	5.03	23.28	0.000
1.1.3 (August)	LARDEN =	3.330 + 6.977 FORAGE2 + 9.081 FORAGE3 + 2.047 FORAGE4 - 0.010 AUG.RAIN	0.60	0.35	5.53	28.89	0.000
1.1.4 (September)	LARDEN =	62.357 + 6.989 FORAGE2 + 9.091 FORAGE3 + 2.068 FORAGE4 - 0.493 SEP.HUMID - 0.994 SEP.TMAX + 0.379 SEP.TMIN + 0.041 SEP.SUNH + 0.005 SEP.RAIN	0.67	0.44	5.10	20.91	0.000
1.1.5 (October)	LARDEN =	77.596 + 6.293 FORAGE2 + 8.707 FORAGE3 + 2.037 FORAGE4 - 0.530 OCT.HUMID - 0.558 OCT.TMAX + 0.647 OCT.TMIN - 1.052 OCT.TAVER + 0.002 OCT.RAIN	0.69	0.48	4.83	22.17	0.000
1.1.6 (November)	LARDEN =	52.954 + 6.989 FORAGE2 + 9.091 FORAGE3 + 2.068 FORAGE4 - 0.427 NOV.HUMID - 0.905 NOV.TMAX +0.664 NOV.TAVER	0.67	0.44	5.07	28.22	0.000

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

The highest value of correlation coefficient (R) was about 0.69 which resulted from the model developed in October (model 1.1.5) and the lowest value of R was about 0.60 from the model developed in August (model 1.1.3). Similarly, the highest and lowest values of coefficient of determination (R^2) were 48% and 35% found from the models developed in October and August respectively. It means that the difference of R^2 between the model developed in October and in August was 13% so the model was developed in October can explain more than 13% of the case comparing with the model develop in August. Other models have the R and R^2 were about 0.67 and 44% respectively. In term of standard error of estimate, it can be seen from the Table 6.1 that the lowest value was 4.83 for the model 1.1.5, and the highest value was 5.53 for the model 1.1.3. It means that the error was higher when the model 1.1.3 was used. However, the difference between them was not significant.

It can be concluded that out of models developed by considering all weather factors and forest age level in each month, the model 1.1.5 which was developed in October was the most appropriate model because its correlation and determination coefficients were highest, and the standard error of estimate of this model was lowest. This model was selected to validate and test in next step.

6.1.2 The models developed by using same weather factor in all months

These models were developed to identify which weather factors had the most significant affect on the density of larvae. The models were set up by testing each individual weather factor in all selected months. Seven models were developed by each weather factor during the period from June to November (Table 6. 2).

The regression coefficients used in the model were significant if their levels are at least 5% by using the t test. The weather factors were tested in different duration of time from June to November. The models set up to experiment average humidity (model 1.2.4) and the model testing the number of sunshine hours (1.2.7) involved only two months. The models set up to evaluate the affect of the number of

rainfall days, involved five months out of the six selected months. However, all of the developed models were significant at the level less than 0.0005 by using the F test. The results of F test are shown in the last two columns of the Table 6.2.

The affect of the forest age was nearly the same as the models developed previously. It means that the larvae density in the forest age level one was the same as that in level five. The larvae densities in other forest age levels are higher than that in the forest age level five. Moreover, the highest larvae density has been found in the forest age level three, followed by the forest age level two, and forest age level four.

The highest value of R was about 0.68 which was resulted from the model developed by the highest temperature (model 1.2.2) and the lowest was about 0.58 for the model developing with number of sunshine hours (model 1.2.7). Similarly, the highest and lowest values of R^2 were 47% and 33% for the models developed by the highest temperature and number of sunshine hours respectively. Moreover, the lowest value of standard of estimate was 5.07 found from the model 1.2.2 (developed by the highest temperature), and the highest value was 5.44 for the model 1.2.3 (developed by the lowest temperature). However, the difference between them was not considerable. The month October was involved in most of the models (six out of seven models), followed by the month September.

It can be concluded that, the model testing the weather factor of the highest temperature was the most appropriate model (model 1.2.2) because it had the highest value of correlation coefficient and coefficient of determination, and moreover it had the lowest value of standard error of estimate. This model was selected to validate and test in next step.

Table 6.2: Regression model for larvae density in December in relation to forest age and each weather factor in all months.

Model	Equation of regression		R	R^2	SE	F	Sig.
1.2.1 (Average temperature)	LARDEN =	6.611 + 6.952 FORAGE2 + 9.298 FORAGE3 + 2.190 FORAGE4 - 1.769 OCT.TAVER - 0.770 AUG.TAVER + 2.243 SEP.TAVER	0.63	0.39	5.40	22.05	0.000
1.2.2 (Highest temperature)	LARDEN =	13.293 + 6.952 FORAGE2 + 9.298 FORAGE3 + 2.190 FORAGE4 + 0.447 AUG.TMAX + 0.420 SEP.TMAX - 0.729 OCT.TMAX - 0.666 NOV.TMAX	0.68	0.47	5.07	25.45	0.000
1.2.3 (Lowest temperature)	LARDEN =	5.241 + 6.952 FORAGE2 + 9.298 FORAGE3 + 2.190 FORAGE4 - 1.013 JUL.TMIN + 0.946 SEP.TMIN + 0.603 OCT.TMIN - 0.788 NOV.TMIN	0.62	0.39	5.44	18.22	0.000
1.2.4 (Humidity)	LARDEN =	14.799 + 6.561 FORAGE2 + 8.915 FORAGE3 + 2.122 FORAGE4 - 0.519 SEP.HUMID + 0.345 OCT.HUMID	0.61	0.37	5.31	23.19	0.000
1.2.5 (Total rainfall)	LARDEN =	0.202 + 6.218 FORAGE2 + 8.910 FORAGE3 + 2.167 FORAGE4 - 0.010 JUN.RAIN - 0.007 AUG.RAIN + 0.004 OCT.RAIN	0.65	0.42	5.14	22.99	0.000
1.2.6 (Number of rainy days)	LARDEN =	- 3.688 + 6.218 FORAGE2 + 8.910 FORAGE3 + 2.167 FORAGE4 - 0.412 JUN.RAIND + 0.439 JUL.RAIND - 0.472 SEP. RAIND + 0.231 OCT.RAIND + 0.406 NOV.RAIND	0.67	0.44	5.08	18.58	0.000
1.2.7 (Number of sunshine hours)	LARDEN =	- 0.613 + 5.776 FORAGE2 + 8.487 FORAGE3 + 2.092 FORAGE4 - 0.028 JUN.SUNH + 0.033 JUL.SUNH	0.58	0.33	5.30	18.31	0.000

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

6.1.3 The model Developed based on selected independent variables

Based on the results of 13 previous models, the independent variables, which were significant at least level 5%, were selected to develop another model. This model was shown in Table 6.3. The value of R, and R^2 of the model was highest. Moreover, its standard error of estimate was lowest. It means that this model can be considered the best model among the models set up by using the multiple linear regression. Besides the model had a significant relationship with the forest age levels. As the same time, this model involved five months out of six months, except July, and it also including six weather factors out of seven factors, except the factor of number of rainfall days. However, the different months or weather factors had the different affect on the larvae density due to different values of regression coefficients. The highest value of regression coefficient was about 1.619 for variable that was the average temperature in June but it was a negative value so when this factor (average temperature in June) increases the larvae density may decrease faster than other factors. The lowest value of regression coefficient was about 0.003 for the variables that were total rainfall in September or October but these regression coefficients were positive values. It means that if these factors (total rainfall in October or September) increase the larvae density may increase. However, affect of these factors on the increase of the larvae density was smaller than those of other factors. The highest temperature involved in three months and the affects were different in different months. For August and September, regression coefficients were positive but it was a negative for November although the measures of affects were nearly the same.

The difference between this model (model 1.3) and previous models was that it involved only one weather factor (total rainfall) in October while other models involved more factors. This may be explained it might an important factor. It may be caused by all selected factors were entered in the model so it can remove some factors, those were less significant, but it may be an important factor. Therefore it is necessary to validate and test these models in next step to select the most appropriate model to predict the larvae density of pine caterpillar in December.

Table 6.3: Regression model to predict the larvae density in December.

Model	Equation of regression	R	R^2	SE	F	Sig.
1.3	LARDEN = 71.161	0.78	0.61	4.15	22.69	0.000
	+ 5.776 FORAGE2					
	+ 8.487 FORAGE3					
	+ 2.092 FORAGE4					
	- 0.820 JUN.HUMID					
	- 0.030 JUN.SUNH					
	- 1.619 JUN.TAVER					
	+ 0.712 AUG.TMAX					
	+ 0.493 SEP.TMIN					
	+ 0.750 SEP.TMAX					
	+ 0.003 SEP.RAIN					
	+ 0.003 OCT.RAIN					
	- 0.692 NOV.TMAX					

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

6.2 The binary logistic regression of epidemic occurrence in December

6.2.1 The models developed by considering all weather factors in each month

Six models were developed by considering all weather factors in each month and the forest age levels. The results of these models are shown in Table 6.4. The forward stepwise conditional method was used to develop these models so it was not necessary to test the significance of the each regression coefficient. By using this method, if the regression coefficient was not significant at level 5%, it was automatically removed from the model.

Table 6.4: Binary logistic regression for epidemic occur in December with all weather factors in each month.

Model	Equation of regression	R^2_N	χ^2	df	Sig.
2.1.1 (June)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = -10.203 + 6.419 FORAGE (1) + 9.510 FORAGE (2) + 10.336 FORAGE (3) + 6.419 FORAGE (4)	0.46	74.98	4	0.000
2.1.2 (July)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = + 0.534 + 6.347 FORAGE (1) + 9.785 FORAGE (2) + 10.860 FORAGE (3) + 6.347 FORAGE (4) - 0.203 JUL.HUMID + 0.016 JUL.RAIN + 0.012 JUL.SUNH	0.56	92.10	7	0.000
2.1.3 (August)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = - 9.796 + 8.121 FORAGE (1) + 10.727 FORAGE (2) + 11.412 FORAGE (3) + 0.000 FORAGE (4) - 0.007 AUG.RAIN	0.54	85.12	5	0.000
2.1.4 (September)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = + 16.031 + 8.087 FORAGE (1) + 10.906 FORAGE (2) + 11.611 FORAGE (3) + 0.000 FORAGE (4) - 0.388 SEP.TMAX - 0.194 SEP.HUMID + 0.018 SEP.SUNH	0.57	94.90	7	0.000
2.1.5 (October)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = + 28.000 + 8.313 FORAGE (1) + 11.492 FORAGE (2) + 12.568 FORAGE (3) + 0.000 FORAGE (4) - 0.446 OCT.TMAX - 0.305 OCT.HUMID + 0.001 OCT.RAIN	0.60	92.89	7	0.000
2.1.6 (November)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = - 9.272 + 8.269 FORAGE (1) + 11.483 FORAGE (2) + 12.182 FORAGE (3) + 0.000 FORAGE (4) + 0.595 NOV.TAVER - 0.562 NOV.TMAX	0.58	97.37	6	0.000

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

The regression coefficients used in the models were significant level at least 5% by using the Wald test. Although some model did not involved the weather factors but the chi-square tests, which is similar to the F test in multiple linear regression, were displayed the significance due to the significance of the forest age levels in the models. The model, which developed in June, had only variable that was the forest age while other weather factors variables were not significant at level 5% under this model. The models developed in July or September or October, on the other hand, involved three weather factors besides the forest age levels.

The forest age level five was used as the reference category. Probability of outbreak occurrence in other forest age levels would be higher than that in level five since all the regression coefficients of these levels were positive. The highest probability was resulted from the forest age level three followed by level two, and level one. Especially, the probability of occurrences in the forest age level four in June and July were higher than level five (reference) but in other months the probability of occurrence in forest age level four and level five was the same because the regression coefficient was equal to zero.

From the Nagelkerke R-square (R^2_N) values, the highest value of R^2_N was about 0.60 for the model developed in October, followed by model in November and September. This result suggested to select the model developed in October to test and validate in next step. However, we can see the goodness-of-fit of all models by using the classification table and the Hosmer and Lemeshow chi-square test that were shown in the Table 6.5.

Table 6.5: Classification table for observed and predicted of occurrence and Hosmer and Lemeshow Test for goodness-of-fit for models developed by each month.

Model	Observed	Predicted		Percentage	Hosmer and Lemeshow Test		
		No occur	Occur	Correct	χ^2	df	Sig.
	No occur	163	21	88.6			
2.1.1	Occur	17	24	58.5			
	Overall			83.1	0.002	2	0.99
	No occur	170	11	93.9			
2.1.2	Occur	15	24	61.5			
	Overall			88.2	6.54	8	0.59
	No occur	166	11	93.8			
2.1.3	Occur	17	21	55.3			
	Overall			87.0	1.89	8	0.98
	No occur	168	12	93.3			
2.1.4	Occur	16	24	60.0			
	Overall			87.3	10.47	8	0.23
	No occur	163	6	96.4			
2.1.5	Occur	17	19	52.8			
	Overall			88.8	6.77	8	0.56
	No occur	166	14	92.2			
2.1.6	Occur	20	20	50.0			
	Overall			84.5	2.61	8	0.96

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

Table 6.5 shows that the highest percentage of correctness was about 88.8% for the model developed in October, followed by model in July (88.2%), and September (87.3%). The lowest percentage of correctness was about 83.1% for the model developed in June. Another thing was that the percentage of correctness might be affected by the difference in sample size. We can see that the highest percentage of correctness in the epidemic none-occurrence was about 96.4% for the model 2.1.5 (developed in October) but this model the percentage correct for epidemic occurrence of this model was too low (about 52.8%) compared with other models, while the

highest percentage of correctness for the epidemic occurrence was about 61.5% for model 2.1.2 (developed in July). Therefore, it was necessary to test and validate both models (developed in July and in October) for choosing the most appropriate model. However, all the models were appropriate since the values of chi-square were small so we cannot reject the hypothesis, there is no difference between the observed and predicted values.

6.2.2 The models developed by using same weather factor in all months

Seven models based on the seven weather factors were developed and the results of these models are shown in the Table 6.6. The results were quite different from those using the multiple linear regression for the larvae density. The models involved very few months. It can be seen from the table that the model for average temperature and the model for the lowest temperature involved only the forest age, and all months were removed from these models. The model for number of rainfall days, however, involved three months, while the models for the humidity and number of sunshine hours involved only one month (September). For forest age variable, the forest level five was used as reference category. The affects by the forest age level four and level five were the same since all regression coefficients of this variable in these models were equal to zero. Like the previous results, the highest occurrence probability resulted from the forest age level three, and followed by forest age level two due to the regression coefficients were highest. Although some model did not involved any months but the chi-square tests displayed the significance that was caused by the significance of the forest age levels in the models.

Table 6.6: Binary logistic regression for epidemic occur in December with each weather factor in all months.

Model	Equation of regression	R^2_N	χ^2	df	Sig.
2.2.1 (Average temperature)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = -10.203 + 6.489 FORAGE (1) + 9.510 FORAGE (2) + 10.203 FORAGE (3) + 0.000 FORAGE (4)	0.48	71.27	4	0.000
2.2.2 Highest temperature)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = -10.481 + 7.592 FORAGE (1) + 11.737 FORAGE (2) + 12.624 FORAGE (3) + 0.000 FORAGE (4) + 0.306 AUG.TMAX - 0.441 NOV.TMAX	0.61	94.86	6	0.000
2.2.3 (Lowest temperature)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = -10.203 + 6.489 FORAGE (1) + 9.510 FORAGE (2) + 10.203 FORAGE (3) + 0.000 FORAGE (4)	0.48	71.27	4	0.000
2.2.4 (Humidity)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = + 7.339 + 7.472 FORAGE (1) + 10.586 FORAGE (2) + 11.384 FORAGE (3) + 0.000 FORAGE (4) - 0.215 SEP.HUMID	0.52	75.54	5	0.000
2.2.5 (Total rainfall)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = - 11.79 + 7.449 FORAGE (1) + 10.825 FORAGE (2) + 12.302 FORAGE (3) + 0.000 FORAGE (4) - 0.009 AUG.RAIN + 0.002 OCT.RAIN	0.65	94.94	6	0.000
2.2.6 (Number of rainy days)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = - 11.109 + 7.462 FORAGE (1) + 10.615 FORAGE (2) + 11.928 FORAGE (3) + 0.000 FORAGE (4) + 0.204 JUL.RAIND - 0.403 SEP.RAIND + 0.216 NOV.RAIND	0.60	85.94	7	0.000
2.2.7 (Number of sunshine hours)	$Ln\left(\frac{P_i}{1-P_i}\right)$ = - 14.122 + 7.534 FORAGE (1) + 10.261 FORAGE (2) + 11.446 FORAGE (3) + 0.000 FORAGE (4) + 0.019 SEP.SUNH	0.53	70.50	5	0.000

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

The highest value of the Nagelkerke R-square (R^2_N) was about 0.65 for the model 2.2.5 which developed by total rainfall, followed by model 2.2.2 with highest temperature (0.61). The lowest value of R^2_N was found from model 2.2.1 (developed by average temperature) and model 2.2.3 (developed by lowest temperature) that was 0.48. These models had the lowest value of R^2 because they involved only the forest age variable. The model 2.2.5 (developed by total rainfall), which had the highest value of R^2, was selected to test and validate in next step.

Besides the Nagelkerke R-square values, we need to consider to the goodness-of-fit that were tested by using the classification table and Hosmer and Lemeshow Test. The results of these tests were shown in Table 6.7. The highest percentage of correctness was about 90.0% for the model developed by number of sunshine hours (2.2.7), followed by model 2.2.5 and model 2.2.6 with 89.2% for testing total rainfall and number of rainfall days. The lowest percentage of correctness was about 82.9% for the model 2.2.1 and model 2.2.3 those were developed by average temperature and the lowest temperature respectively. Another thing is that the percentage of correctness might be affected due to the difference in the sample size. It can be seen from the table that the highest percentage of correctness for predicting the epidemic none-occurrence was about 96.9% for the model 2.2.7 (developed by number of sunshine hours) but the percentage of correctness for predicting the epidemic occurrence of this model was too low (about 53.3%) compared with other models, while the highest percentage of correctness for predicting the epidemic occurrence was about 65.6% for model 2.2.5 (developed by total rainfall). It can be concluded that the model 2.2.5 was the best model among the seven models developed by considering each weather factor and forest age. However, all the models were appropriate since the values of chi-square were small therefore we could not reject the hypothesis, there is no difference between the observed and predicted values.

Table 6.7: Classification table for observed and predicted of occurrence and Hosmer and Lemeshow Test for goodness-of-fit for models developed by each factor.

Model	Observed	Predicted		Percentage	Hosmer and Lemeshow Test		
		No occur	Occur	Correct	χ^2	df	Sig.
	No occur	153	21	87.9			
2.2.1	Occur	15	21	58.3			
	Overall			82.9	0.003	3	1.00
	No occur	165	9	94.8			
2.2.2	Occur	15	21	58.3			
	Overall			88.6	3.59	8	0.89
	No occur	153	21	87.9			
2.2.3	Occur	15	21	58.3			
	Overall			82.9	0.003	3	1.00
	No occur	162	9	94.7			
2.2.4	Occur	18	16	47.1			
	Overall			86.8	3.37	8	0.91
	No occur	153	10	93.9			
2.2.5	Occur	11	21	65.6			
	Overall			89.2	1.59	8	0.99
	No occur	154	9	94.5			
2.2.6	Occur	12	20	62.5			
	Overall			89.2	6.73	8	0.57
	No occur	155	5	96.9			
2.2.7	Occur	14	16	53.3			
	Overall			90.0	2.45	8	0.96

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

6.2.3 The model based on the selected independent variables

Based on the results of 13 previous models, the independent variables, which were significant at 5% level, were selected to develop another model. This model was shown in Table 6.8. Furthermore, the classification table and the Hosmer and Lemeshow Test of this model were presented in the Table 6.9.

Table 6.8: Binary logistic regression for epidemic occur in December.

Model	Equation of regression	R^2_N	χ^2	df	Sig.
2.3	$Ln\left(\frac{P_i}{1-P_i}\right) =$ + 14.276 + 8.924 FORAGE (1) + 14.804 FORAGE (2) + 17.036 FORAGE (3) + 0.000 FORAGE (4) + 0.033 JUN.SUNH + 0.022 JUL.RAIN - 0.010 AUG.RAIN - 0.463 SEP.HUMID + 0.002 OCT.RAIN	0.78	114.53	9	0.000

(Source: TTH DFP, 2004 and Analyzed by SPSS 10.0).

The Nagelkerke R-square value of the model was higher than those of previous models. It was about 0.78. Similar to the previous models, the forest age was the significant variable and the forest age level five was used as a reference category. The model involved five months, but it involved only three weather factors namely, total rainfall, number of sunshine hours, and average humidity. However, the probability of the epidemic occurrence resulted from the different months and different weather factors were different due to different values of regression coefficients. The highest value of regression coefficient was about 0.463 for the average humidity in September variable. However, it was a negative value so when this factor (humidity in September) increases the probability of epidemic occurrence decrease and the decrease caused by the factor might happen faster than other weather

factors. The lowest value of regression coefficient was about 0.002 for the total rainfall in October variable and it was a positive value, therefore the probability of epidemic occurrence will increase when this factor (total rainfall in October) increases but the its effect was be smaller than other factors. The total rainfall was involved in three months and its effects varied with months. In July and October, it was positive but negative in August.

Besides the Nagelkerke R-square values, we need to consider to the goodness-of-fit that were tested by using the Hosmer and Lemeshow Test. The table 6.9 showed that a high overall of about 94.2%. Whereas, the correctness for predicting the none-occurrence and occurrence situation was about 96.3% and 83.3%, respectively. It can be concluded that the model developed by using the selected independent variables was the best model compared with other binary logistic regression models. However, this model needs to be tested and validated before using and making a conclusion.

Table 6.9: Classification table for observed and predicted of occurrence and Hosmer and Lemeshow Test for goodness-of-fit for models developed by selected variables.

Model	Observed	Predicted		Percentage	Hosmer and Lemeshow Test		
		No occur	Occur	Correct	χ^2	df	Sig.
	No occur	154	6	96.3			
2.3.	Occur	5	25	83.3			
	Overall			94.2	0.53	8	1.00

(Source: TTH-DFP, 2004 and Analyzed by SPSS 10.0).

6.3 Model validation

6.3.1 Validating the multiple linear regression

Three models were selected to test and validate the larvae density in December including model 1.1.5, model 1.2.2, and model 1.3. These models were used the method of multiple linear regression and developed by considering all factors in October, by testing the highest temperature in all months, and by using selected independent variables respectively. The paired samples t-test and Wilcoxon signed-ranks test were used to test the models. In addition, the root mean square error (RMSE) was calculated to identify the appropriateness of the models. These results are shown in Table 6.10.

Table 6.10: Validation test of larvae density of three selected regression models.

Testing methods	Model 1.1.5	Model 1.2.2	Model 1.3
Paired samples t-test			
Sample size	91	96	91
Mean	3.67	4.72	3.76
Different mean	0.401	- 0.403	0.303
t value	0.704	- 0.752	0.665
Sig.	0.483	0.454	0.508
Wilcoxon signed-ranks test			
Sample size	91	96	91
Sum of negative ranks	1896	1771	2061
Sum of positive ranks	2290	2885	2125
Z value	- 0.780	- 2.035	-0. 127
Sig.	0.436	0.042	0.899
RMSE	5.418	5.239	4.338

(Source: Field survey, 2004 and Analyzed by SPSS 10.0).

The sample used for testing the models consisted of 85 cases in 17 years (from 1987 to 2003) recorded in other district, Huong Tra district, and another 16 cases were

collected during the field survey, in December 2004, from four districts and in four levels of forest age. However, some cases were missing when testing due to missing some value of weather factors.

Table 6.10 shows that all of the three models were appropriate by using the paired samples t-test as well as by Wilcoxon signed-ranks test since all models were not significant at 5% level, except model 1.2.2 in Wilcoxon signed ranks test of which the significance was 0.042. By using the two testing methods, the highest significance was for model 1.3, followed by model 1.1.5, and model 1.2.2. However, if RMSE was considered, the lowest value of RMSE was found from the model 1.3, followed by model 1.2.2. The t values in paired sample t-test were positive for model 1.1.5 and 1.3 but it was negative for model 1.2.2. It means that the model 1.2.2 was an overestimated model while model 1.1.5 and model 1.3 were underestimated models. All the Z values in Wilcoxon signed-ranks test were negative showing that the sum of negative ranks was less than the mean, which had smaller value than the sum of positive ranks.

From those indicators, it can be concluded that among the developed models the model 1.3, which developed by using the forest age, average humidity in June, number of sunshine hours in June, average temperature in June, highest temperature in August, September and November, lowest temperature in September, and total rainfall in September and October, was the most appropriate model to predict the larvae density in December in Thua Thien Hue province. It is strongly recommended to use model 1.3, which using forest age and weather variables from different months, to predict the larvae density in this province because it had the highest value of R or R^2, and lowest standard error of estimate. Moreover, in validation test, it provided a non-significant difference between the observed and predicted larvae density in December by using paired samples t-test and Wilcoxon signed-ranks test, and has the lowest value of RMSE.

6.3.2 Validating the binary logistic regression

For the probability of epidemic occurrence, four models have been tested and validated to select to most appropriate one. They were model 2.1.2, model 2.1.5, model 2.2.5, and 2.3 which were developed by considering all factors in July, in October, developed by testing total rainfall in all months, and by using selected independent variables respectively. The classification table and symmetric test (Phi indicators) were used to select the appropriate model. The results were shown in the Table 6.12.

It can be seen from table that the highest value of percentage of correctness was about 96% for predicting the epidemic none-occurrence in model 2.1.5. This model also had the lowest value of percentage of correctness for predicting the epidemic occurrence with about 20%. The model 2.3 had the lowest overall percentage of correctness about 81% while this value in the model 2.2.5 was about 86%. However, these models were used to predict the epidemic occurrence so the percentage of correctness for predicting epidemic occurrence should be as higher as possible. In term of this, the model 2.3 was the best one because its percentage of correctness was about 62%, followed by model 2.2.5 with about 60%, while model 2.1.5 has only 20%.

To test the symmetric for the nominal-by-nominal variables the Phi indicator was used. The highest value of Phi was about 0.495 for the model 2.2.5, which developed by total rainfall in August and October, followed by model based on the selected independent variables of about 0.401. Whilst the lowest of Phi value was about 0.24 for model 2.1.5, which developed by using highest temperature, humidity, and total rainfall in October. However, all the models were significant at 5% level.

By combining both indicators, classification table and Phi test, the model 2.2.5 should be the most appropriate model, followed by model 2.3 among the developed models. Moreover, the difference of both indicators between model 2.2.5 and model 2.3 was not much and was not significant. Therefore, the model 2.2.5 and model 2.3

can be used to predict the pine caterpillar epidemic occurrence without significant difference.

Table 6.11: The classification table and symmetric test for binary logistic models.

Model	Prediction	Observation in December			Percentage	Symmetric test	
		No occur	Occur	Total	correct	Phi value	Sig.
	No occur	73	7	80	91.3		
2.1.2	Occur	9	7	16	43.8	0.370	0.000
	Total	82	14	96	83.3		
	No occur	73	3	76	96.1		
2.1.5	Occur	12	3	15	20.0	0.240	0.022
	Total	85	6	91	83.5		
	No occur	69	7	76	90.8		
2.2.5	Occur	6	9	15	60.0	0.495	0.000
	Total	75	16	91	85.7		
	No occur	62	11	73	84.9		
2.3	Occur	5	8	13	61.5	0.401	0.000
	Total	67	19	86	81.4		

(Source: Field survey, 2004 and Analyzed by SPSS 10.0).

It can be concluded that statistical models can be developed to predict not only larvae density but also probability of epidemic occurrence. The larvae density in December can be derived by using the model 1.3 that was developed by using multiple linear regression with forest age, humidity in June, Sunshine duration in June, averaged air temperature in June, maximum air temperature in August, September, and November, minimum air temperature in September, and total rainfall in September and October. The probability of epidemic occurrence can be derived from the model 2.2.5 that was developed by using binary logistic regression with total rainfall in August and October or from model 2.3 that using binary logistic regression based on forest age, total rainfall in July, August and October, number of sunshine hours in June, and humidity in September.

Despite the fact that these models were the most appropriate models because they had high the correlation coefficients, and low errors, they need to be remodeled and tested with other factors because the coefficient of determinations were still not high enough and the RMSE were still high. It means that the models cannot explain exactly all of the cases, and it also means that the models did not include enough independent variables to explain the real world. The limitation of this study were the data that were not collected enough in all months, and very few variables were selected so the developed models might not satisfy all the cases. In addition, it is necessary to refine the binary logistic regression model or to add to the models some more independent variables such as level of natural enemies, biomass of pine-leaves (food sources), or daily weather factors because the percentage of correctness was not fairly high, especially the percentage of correct for predicting the epidemic occurrence.