Distribution of Horseshoe Crabs Satellites Analyzed by Logistic Regression

Ouhang Gong

Biological Engineering faculty, East China University of Science and Technology, Shanghai 200237, China

2008120109@st.btbu.edu.cn

Keywords: horseshoe crabs, satellite, logistic regression.

Abstract: Pairs of horseshoe crabs arrive at the beach and spawn during spring. Single males go to the beach as well, and they make competition against males with partners for fertilization. The group size of satellite males is so large that there is a nonrandom distribution in which habitat selection and local environmental conditions are not responsible. The experimental manipulations show that satellites cannot only copy other male satellites' behaviors. According to the evidence revealed by experiments and observations, the possible explanation about the nonrandom distribution of satellite males located among nesting pairs is that it is more likely for unattached satellite males to attract females than others. To analyze the data from this study, R was used to generate the confusion matrix, which checks if the predictable value is close to the actual value, logistic regression, which checked if the model should be modified, and logit function to evaluate the model. As a result, logistic regression showed that the model should be modified because it was not proper. Width and weight were selected as variables. In conclusion, the logistic regression showed that a male horseshoe crab with larger width and weight was more likely to own satellites. Also, the color of a horseshoe crab was related to the satellite.

1. Introduction

Horseshoe crabs are marine and brackish water arthropods of the family Limulidae and the only living members of the order Xiphosura [2]. The horseshoe crab is under State second-class protection, and the study of their breed may help them not to go extinct [3]. What is more, the blood of the horseshoe crabs can be applied in the detection of bacteria and toxin contamination [4, 5]. Therefore, the study of horseshoe crabs has important significance for both humans and the environment.

Pairs of horseshoe crabs arrive at the beach and spawn during spring [11]. Single males, which are seen as "Satellites", go to the beach as well, and they make competition against males with partners for fertilization [1]. The group size of satellite males is so large that there is a non-random distribution in which habitat selection and local environmental conditions are not responsible [6, 7]. Interestingly, horseshoe crabs are well known for their mass, springtime spawning and strongly male-biased, operational sex ratios and the "satellite problems" is a very classic statistic problem [8, 10].

H. Jane Brockmann fund Females with satellites laid no more eggs than those found nesting without satellites Group size was also not correlated with a total number of eggs laid. Some couples were more likely than others to attract unattached males as satellites [11]. David R. SMITH used logistic regression to indicate that animal size as measured by weight or prosomal width was not related to the likelihood of future recapture for either sex [8]. Smith David R. used statistical analysis to confirm horseshoe crab's spatial and temporal distribution (Limulus polyphemus) spawning in Delaware Bay [12]. But no article links the "satellite" phenomena and the characteristics of Horseshoe crabs together.

This article conducts several analyses to explain the distribution of satellites. Section 2.1 introduces the data. Section 2.3 describes the method used in the study. Section 3 talks about the result.

2. Method

2.1 Data collection

The data was collected by the team of Brockmann, Monica Marquiiz, Alfonso Aionso-Mej and Carlos Iudica. Susan Wineriter, Eric Botsford and Christian Solare and especially Kim Abplanalp for assisting with the field work. The University of Delaware, College of Marine Studies, Lewes, DE provided research facilities and accommodation, and the Cape Henlopen State Park permitted to work on the Breakwater Harbor beach. This research was supported by the National Science Foundation OCE 90-06392 and by the Florida Foundation [11].

The data was collected at two beaches on the Delaware shore, Breakwater Harbor at Cape Henlopen State Park in Lewes and Fowler's Beach, 32 km north on the same shoreline (Sussex County, Delaware, USA). In 1991 observations were made from 7 to 17 June, in 1992 from 28 May to 3 June and from 1 1 to 14 June, and in 1993 from 18 May to 11 June. At these sites, the crabs were most active on the higher of the two daily high tides (which at this time of year are at night between 1700 and 0200 h EST) [11]. In the data, 173 horseshoe crabs were tested. The number of "satellites" and whether they have "satellite" were considered dependent variables, and the characteristics, which are their weight, width, colour, and the spine, were considered independent variables.

The raw data was accessed from this address: http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat. And the data analysis tool was R studio i386 4.1.0.

2.2 Research Variables

There were 173 entries in the data. There were 2 kinds of dependent variables: numerical variable, which was the number of satellites and category variable, which was whether crabs have satellites. There were 4 independent variables which were weight, width, colour and spine. Variable colour means the colour of the horseshoe crabs. The number of colours means the degree of colour depth. The bigger number, the darker colour. Variable spine means the number of the spine that horseshoe crabs have.

In the regression, responsive variable y was whether the female has satellites. y equalled 1 when there were satellites, and y equals 0 when there were not.

Table 1. The basic information of the data.

X	**Type**	**Range**	**Amount**
Weight	numerical	[1.200,5.200]	
Width	numerical	[21.0,33.5]	
Color	categorical	{1,2,3,4}	12 color=1(6.94%) 95 color=2(54.91%) 44 color=3(25.43%) 22 color=4(12.72%)
spine	numerical	{1,2,3}	37 spine=1(21.39%) 15 spine=2(8.67%) 121spine=3(69.94%)
Y	Type	Range	Amount
y	categorical	{0,1}	62 y=0 (35.84%) 111 y=1 (64.16%)
sat	numerical	{0,1,2,3,4,5,6,7,8,9,10,11,12,14,15}	

2.3 Model selection and Statistical analysis

First, we preprocess the data. We create the new data frame (weight, width, color, spine, y) and with y in the last column without sitting. The new data has one dependent variable and four independent variables. Second, we fit the model (1). However, none of the effects is significant except the Intercept, which indicates that this model is not proper, and modification is needed. Table 2 also indicates that model (1) does not fit it very well.

Table 2. Coefficients of the model (1)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.06501	3.92855	-2.053	0.0401 *
weight	0.82578	0.70383	1.173	0.2407
width	0.26313	0.19530	1.347	0.1779
factor(color)2	-0.10290	0.78259	-0.131	0.8954
factor(color)3	-0.48886	0.85312	-0.573	0.5666
factor(color)4	-1.60867	0.93553	-1.720	0.0855 ·
factor(spine)2	-0.09598	0.70337	-0.136	0.8915
factor(spine)3	0.40029	0.50270	0.796	0.4259

As the value of weight increases, the value of width increases as well. There are two dependent variables which are sat and y. We are interested in weight as a coefficient for the dependent variable sat because it is the most significant one with the significant code of 0.05. But other coefficients lack significant codes. This reason can be applied to the dependent variable y as well. For this one, we are interested in coefficient color since its significant code is 0.01, which means that this is the most significant variable. Other coefficients don't have significant code.

And there seems multicollinearity between weight and width, which lead to the insignificance of both effects in original, and they split the explanatory effect evenly.

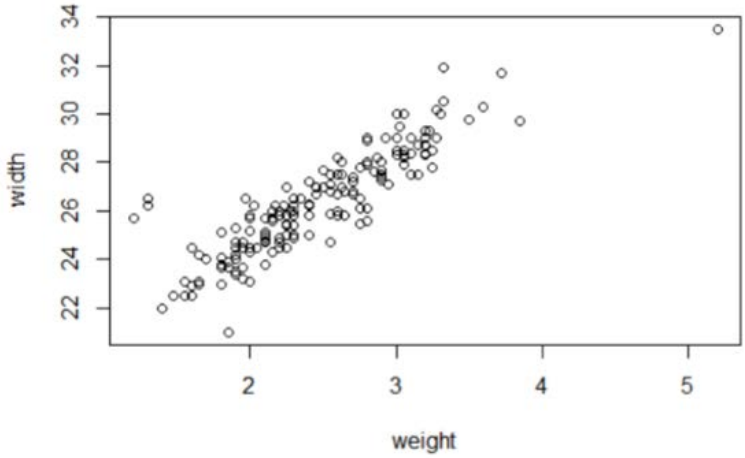

Figure 1. The multicollinearity between weight and width

Therefore, in the final model, we only need the better one of the weight and width. In the further exploration of the model, we try with the methods Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [15,16]. We use AIC as a reference to make a selection. After selection, in the new model, the remaining explanatory variables are width and colour.

2.4 Strategies in Model Selection

There were 2 continuous variables which are weight and width. Mean Squared Error (MSE) was used to check if actual values are close to predictable values [17]. The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections, the formatting shown in Table 2 should be used.

There were two categorical variables: color and spine, so logistic regression was used to get a confusion matrix [18, 19]. The coefficient of Variation (CV) method was used to select lambda and fit the best lam [20]. Then we tested the confusion matrix.

2.5 Model Evaluation

Since we got a logic model, logic function was used: $P(Y=1 \mid X=x)=\exp(x'\beta)/(1+\exp(x'\beta))$ where P was the probabilities of the outcome to evaluate [13]. And the cook's distance was used to test the outliers [14].

3. Result

The regression result was:

$$logit(y) = -8.07 + 0.83weight + 0.26width - 0.10c2 - 0.49c3 - 1.61c4 - 0.10s2 - 0.40s3 \quad (1)$$

$$logit(y) = -11.38 + 0.47width + 0.07c2 - 0.229c3 - 1.33c4 \quad (2)$$

Table 3. Coefficients of the two models

	r	weight	width	c2	c3	c4	s2	s3
Logit(1)	-8.07	0.83	0.26	-0.10	-0.49	-1.61	-0.10	-0.40
Logit(2)	-11.38	0.47	/	0.07	-0.229	-1.33	/	/

3.1 Logistic Regression

In model (1), c meant color and s meant the number of spines. The constant was -8.07, the coefficient of weight was +0.83, the coefficient of width was +0.26, the coefficient of c2 was -0.10, the coefficient of c3 was -0.49, the coefficient of c4 was -1.61, the coefficient of s2 is -0.10, the coefficient of s3 is -0.40. However, none of the effects was significant except the Intercept, which indicated that this model was not proper, and modification was needed.

3.2 Variable Selection

The multicollinearity between weight and width split the explanatory effect evenly. So, in the final model (2), a better one of the width and weight was needed. Akaike Information Criterion (AIC) was used as a reference to make a selection [16]. The AIC value of the original model (1) was 201.2.

For y, Mean Squared Error (MSE) was associated with width of 0.3006361, and MSE was associated with a weight of 0.3046534. Because these two values were small, the estimated weight and width affecting y were close to the actual value. For sat, MSE associated with width was 13.58479, and MSE associated with weight was 13.26504. Therefore, the value of MSE was large, so the estimated value of weight and width affecting sat was not close to the actual value. The confusion matrix showed that there were 27 true negatives and 16 false negatives. The per cent of correct classification is 69.77%, which is a rather good result.

Bayesian Information Criterion (BIC) was used to choose too [15]. The BIC q equivalent for q in (0.477740316103793, 0.876695783647898), the best model was also "width" and "colour". The estimate for width is 0.4583097, and for colour is -0.5090467. The Std. Error of width was 0.1040181, and of colour was 0.2236817. The z value of width was 4.406056, and of colour was -2.275763. The Pr (>|z|) of width was 1.052696e-05 and of colour is 2.286018e-02. So, the variables chosen were the same.

After selection, in the new model (2), the remaining explanatory variables were width and colour, with AIC 197.5. The constant value was -11.38, the coefficient of width was 0.47, the coefficient of c2 was +0.07, the coefficient of c3 was -0.229, the coefficient of c4 was -1.33. When compared with the backward selection y ~ weight + width + factor (colour) + factor(spine), with AIC==201.2, y ~ weight + width + factor (colour), with AIC==198.21, which were all higher than our current model. And for the interaction, when the model was tried to fit with y ~ width +I(width^2) +factor (colour), the value of AIC was 198.6062, and the model of y ~ width +factor (colour)+factor (colour)*width whose AIC was 199.0806, which were both bigger than the first model. So, we will not add interaction to the model. Finally, the model (2):

logit(y) = −11.38 + 0.47width + 0.07c2 − 0.229c3 − 1.33c4 (2) was fitted.

3.3 Model Evaluation

There were three outliers. These three data were removed: the 23rd Crab whose Cook's distance is 1.06e-10; the 72nd Crab whose Cook's distance is 4.39e-11 and the 99th Crab whose Cook's distance is 3.24e-11. And we get new data with 170 entries.

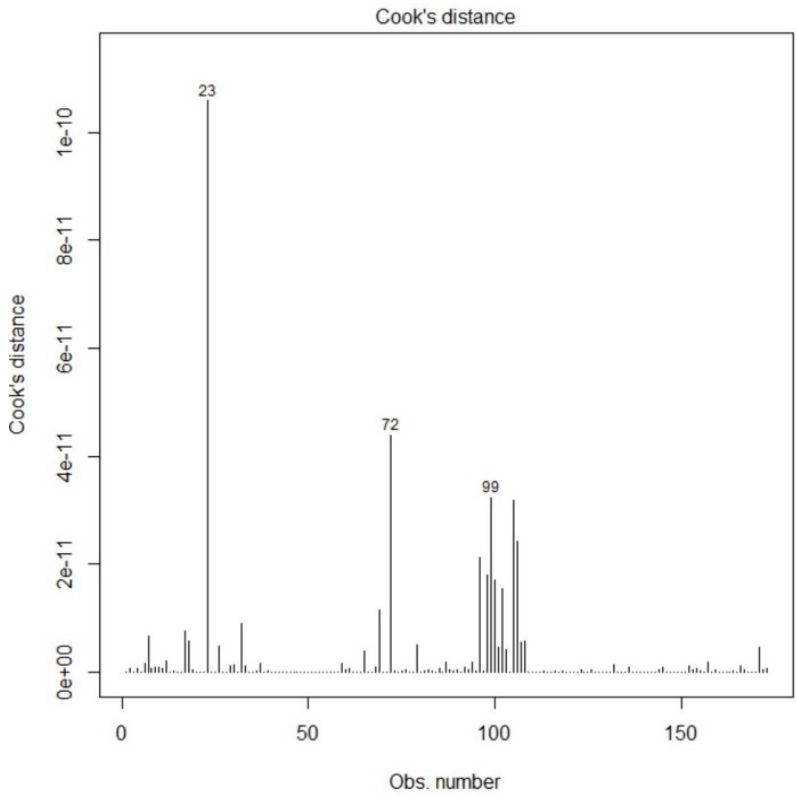

Figure 2. The cook's distance result

For the Variance Inflation Factor (VIF) of the model, sat=1.019763, weight=3.544348, width=3.575887, colour=1.309422, spine=1.290683. All of them were less than 10, so there was no obvious multicollinearity between the independent variables. The result can be seen in Figure 3.

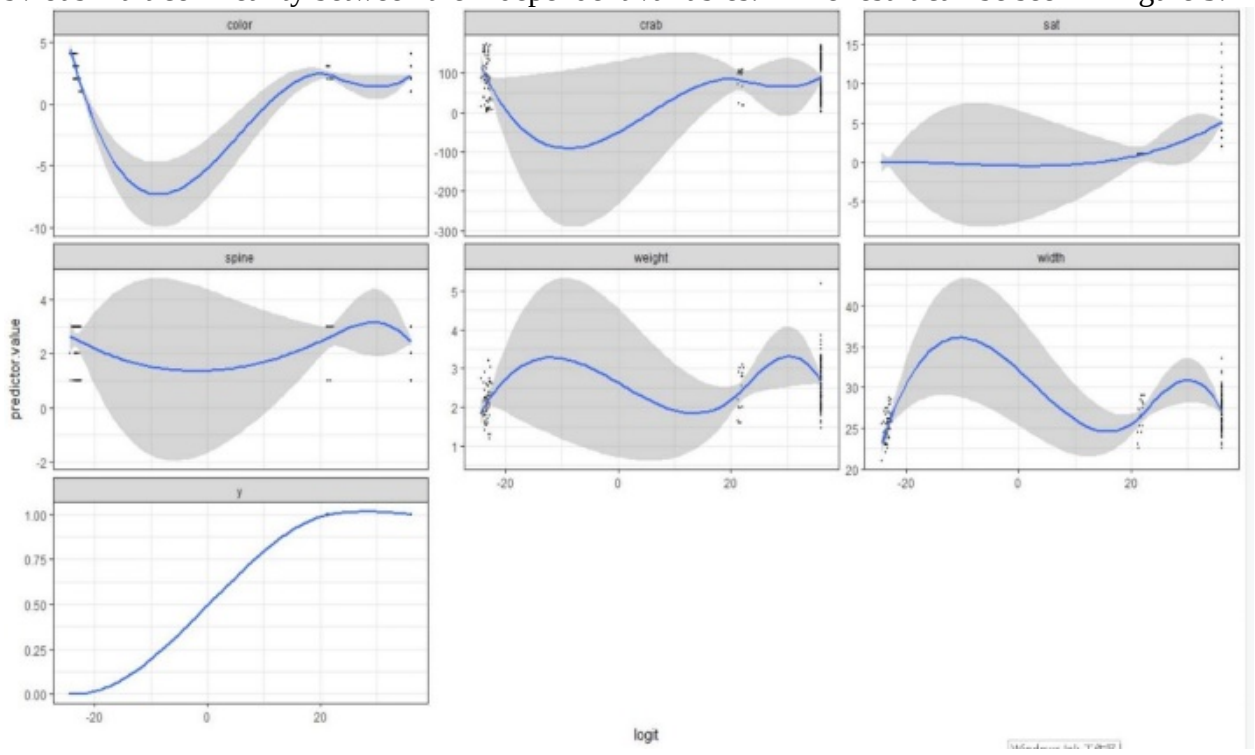

Figure 3. The predictor-value result

Figure 3 shows the predictor value of y and the independent variables. The figure showed that colour and width have a smaller range of predictor values, which also confirmed that our model variables (width and colour) fit better with y.

4. Discussion

The horseshoe crab is under State second-class protection, and the research on their reproduction help them avoid extinction. What is more, the blood of the horseshoe crabs can be applied in the detection of bacteria and toxin contamination. Therefore, the study of horseshoe crabs has important significance for both humans and the environment.

And our model provides a possibility to monitor and predict the number of horseshoe crabs.

But our model has some limitations. First, the data was collected almost nearly 30 years ago, and it may not suit the horseshoe crabs who live in the current environment now. Second, the amount of data is a little bit small (with only 173 entries), which may have a greater occasionality in the result. Third, this is an obvious model for horseshoe crab behavior, but nothing is known, as yet, about the rules that individual males use to join or leave groups. We still need a lot of work to study the habits of horseshoe crabs and get to know more about them. Finally, unattached, satellite and attached males are not equal competitors, but it is not clear how this will change their behavior when they are in such complex, interacting groups. What's more, we only tried to fit several models. There are more possibilities, and our model may not be the best. Further work has to be done to study more about the habits of horseshoe crabs and try to fit a better model.

5. Conclusion

In conclusion, a logistic model was fitted, which was $\text{logit}(y) = -11.38+0.47\text{width}+0.07c2-0.229c3-1.33c4$. logit(y) had a positive correlation to "width". Since "width" and "weight" had collinearity, it could be concluded that a horseshoe crab with bigger weight and wider width had satellites.

Whether a horseshoe crab had satellites had some correlation to their color. A horseshoe crab with color 2 could increase the possibility of a satellite, but a horseshoe crab with color3 or 4 could decrease the possibility of having a satellite. In the future, we want to find out what lead to this result and more quality about the "satellite".

References

[1] H. Jane Brockmann. Mating Behavior of Horseshoe Crabs, Limulus polyphemus [J]. Behaviour, 1990, 114(1/4)

[2] Gorman Richard. Atlantic Horse Crabs and Endotoxin Testing: Perspectives on Alternatives, Sustainable Methods, and the 3Rs (Replacement, Reduction, and Refinement) [J]. Frontiers in Marine Science, 2020,

[3] H. Jane Brocckmann, Colette M. St Mary, José Miguel Ponciano. Discovering structural complexity and its causes: Breeding aggregations in horseshoe crabs [J]. Animal Behaviour, 2018, 143: 177-191.

[4] Tamura Hiroshi, Reich Johannes, Nagaoka Isao. Outstanding Contributions of LAL Technology to Pharmaceutical and Medical Science: Review of Methods, Progress, Challenges, and Future Perspectives in Early Detection and Management of Bacterial Infections and Invasive Fungal Diseases. [J]. Biomedicines, 2021, 9(5).

[5] Mark A. Beekey, Jennifer H. Mattei, Barbara J. Pierce. Horseshoe crab eggs: A rare resource for predators in Long Island Sound [J]. Journal of Experimental Marine Biology and Ecology, 2013, 439.

[6] Tan A N, Christianus A, Shakibazadeh S, Hajeb P. Horseshoe crab, Tachypleus gigas (Müller, 1785) spawning population at Balok Beach, Kuantan, Pahang, Malaysia. [J]. Pakistan journal of biological sciences: PJBS, 2012, 15(13).

[7] S. GILLINGS, P. W. ATKINSON, S. L. BARDSLEY, N. A. CLARK, S. E. LOVE, R. A. ROBINSON, R. A. STILLMAN, R. G. WEBER. Shorebird predation of horseshoe crab eggs in

Delaware Bay: species contrasts and availability constraints [J]. Journal of Animal Ecology, 2007, 76(3).

[8] David R. Smith. Effect of horseshoe crab spawning density on nest disturbance and exhumation of eggs: A simulation study. Estuaries and Coasts, 2007, 30(2).

[9] John A. Sweka, David R. Smith, Michael J. Millard. An age-structured population model for horseshoe crabs in the Delaware Bay area to assess harvest and egg availability for shorebirds [J]. Estuaries and Coasts, 2007, 30(2).

[10] Sara P. Grady,Ivan Valiela. Stage-structured matrix modeling and suggestions for management of atlantic horseshoe crab, Limulus polyphemus, populations on cape cod, Massachusetts [J]. Estuaries and Coasts, 2006, 29(4).

[11] H. Jane Brocckmann, Satellite Male Groups in Horseshoe Crabs, Limulus polyphernus [J]. Ethology 102, 1-21 (1996)

[12] Smith David R., Pooler Penelope S., Swan Benjie L., Michels Stewart F., Hall William R., Himchak Peter J., Millard Michael J.. Spatial and temporal distribution of horseshoe crab (Limulus polyphemus) spawning in Delaware Bay: Implications for monitoring [J]. Estuaries, 2002, 25(1).

[13] Vollmer RT. Multivariate statistical analysis for pathologists. Part I. The logistic model. Am J Clin Pathol 1996; 105: 115-26.

[14] Rius F.X., Smeyers-Verbeke J., Massart D. L.. Method validation: software to plot calibration lines and their response residuals, and to detect outliers according to Cook's distance[J]. Elsevier,1989,8(1).

[15] A. E. Raftery. A Note on Bayes Factors for Log-Linear Contingency Table Models with Vague Prior Information [J]. Journal of the Royal Statistical Society. Series B (Methodological), 1986, 48(2).

[16] Ehlinger Gretchen S., Tankersley Richard A., Bush Mark B.. Spatial and temporal patterns of spawning and larval hatching by the horseshoe crab, Limulus polyphemus, in a microtidal coastal lagoon [J]. Estuaries, 2003, 26 (3).

[17] Wallach D., Goffinet B.. Mean squared error of prediction as a criterion for evaluating and comparing system models [J]. Elsevier, 1989, 44(3-4).

[18] Vollmer RT. Multivariate statistical analysis for pathologists. Part I, The logistic model. Am J Clin Pathol 1996; 105:115–26.

[19] Lemeshow S, Hosmer DW. Logistic regression. In: Armitage P, Col- ton T, Eds. Encyclopedia of Biostatistics. New York: J.Wiley, 1998. p. 2316–27.

[20] Faber D.S., Korn H.. Applicability of the coefficient of variation method for analyzing synaptic plasticity [J]. Cell Press, 1991, 60(5).