

Clasificación

Maestría en Biotecnología

Daniel Martínez Bello

Universidad de Santander

5 de diciembre de 2025

Contenido

- Ejemplo motivador
- Regresión logística
- Actividad

Cuadro: Base de datos, cangrejo herradura

crab	sat	y	weight	width	color	spine
1	8	1	3	28	2	3
2	0	0	2	22	3	3
3	9	1	2	26	1	1
4	0	0	2	25	3	3
5	4	1	3	26	3	3
6	0	0	2	24	2	3
7	0	0	2	26	1	1
8	0	0	2	25	3	2
9	0	0	2	24	2	1
10	0	0	2	26	3	3
11	0	0	2	24	3	3
12	0	0	3	26	2	3
13	11	1	3	28	2	3
14	0	0	2	21	4	2
15	14	1	2	26	2	1
16	8	1	3	27	1	1
17	1	1	2	25	2	3
18	1	1	3	29	2	3
19	0	0	2	25	4	3
20	5	1	3	27	2	3

Información de las variables

- id: identificación del cangrejo
- sat: número de machos adheridos a cada hembra (conteo)
- weight: peso de cada hembra (discreta dicotómica)
- width: ancho del caparazón (continua)
- color: color del caparazón, 1: leve, 2: medio, 3: medio oscuro, 4: oscuro (discreta, ordinal)
- spine: condición de las espinas, 1: ambas buenas, 2: una buena, una dañada, 3: las dos dañadas (discreta ordinal)

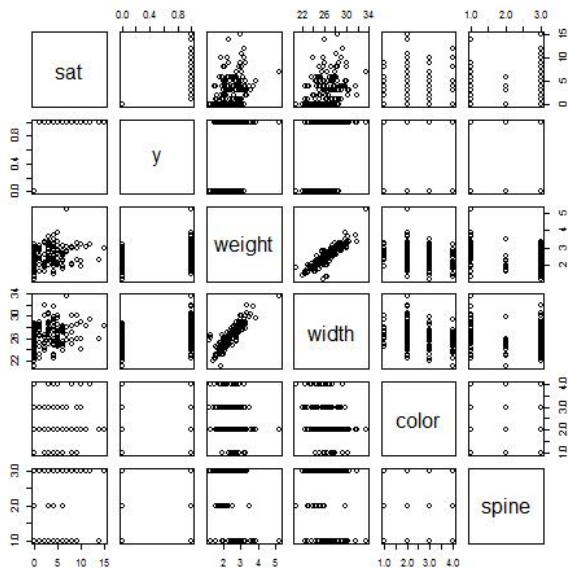
Cangrejo herradura



Cangrejos herradura



Matriz de dispersión



Cuadro: Estadística descriptiva

label	levels	0	1	Total	p
Total N (%)		62 (35.8)	111 (64.2)	173	
sat	Mean (SD)	0.00 (0.00)	4.55 (2.83)	2.92 (3.15)	<0.001
weight	Mean (SD)	2.14 (0.45)	2.60 (0.58)	2.44 (0.58)	<0.001
width	Mean (SD)	25.17 (1.67)	26.93 (2.07)	26.30 (2.11)	<0.001
colorfactor	1	3 (4.8)	9 (8.1)	12 (6.9)	0.003
	2	26 (41.9)	69 (62.2)	95 (54.9)	
	3	18 (29.0)	26 (23.4)	44 (25.4)	
	4	15 (24.2)	7 (6.3)	22 (12.7)	
spinefactor	1	11 (17.7)	26 (23.4)	37 (21.4)	0.272
	2	8 (12.9)	7 (6.3)	15 (8.7)	
	3	43 (69.4)	78 (70.3)	121 (69.9)	

Pregunta problema

¿Como pueden los investigadores de los cangrejos herradura clasificar si una cangrejo hembra es exitosa en tener al menos un cangrejo macho satélite de acuerdo a sus características biométricas?

Regresión logística simple

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i$$
$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Regresión logística múltiple

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$
$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}$$

Estadística descriptiva

Cuadro: Descriptive Statistics

	weight		width		color		spine	
	0	1	0	1	0	1	0	1
Valid	62	111	62	111	62	111	62	111
Mean	2,139	2,604	25,169	26,930	2,726	2,279	2,516	2,468
Std. Deviation	0,448	0,575	1,674	2,069	0,890	0,703	0,784	0,851
Minimum	1,200	1,475	21	22,500	1	1	1	1
Maximum	3,200	5,200	28,700	33,500	4	4	3	3

Cuadro: Model Summary - y

Model	Deviance	AIC	BIC
H ₀	225,759	227,759	230,912
H ₁	185,202	201,202	226,428

Cuadro: Coefficients

	Estimate	Standard Error	z	Wald Test		
				Wald Statistic	df	p
(Intercept)	-8,065	3,929	-2,053	4,214	1	0,040
weight	0,826	0,704	1,173	1,377	1	0,241
width	0,263	0,195	1,347	1,815	1	0,178
color (2)	-0,103	0,783	-0,131	0,017	1	0,895
color (3)	-0,489	0,853	-0,573	0,328	1	0,567
color (4)	-1,609	0,936	-1,720	2,957	1	0,086
spine (2)	-0,096	0,703	-0,136	0,019	1	0,891
spine (3)	0,400	0,503	0,796	0,634	1	0,426

Cuadro: Model Summary - y

Model	Deviance	AIC	BIC
H ₀	225,759	227,759	230,912
H ₁	186,211	198,211	217,131

Cuadro: Coefficients

	Estimate	Standard Error	z	Wald Test		
				Wald Statistic	df	p
(Intercept)	-8,644	3,770	-2,293	5,257	1	0,022
weight	0,773	0,698	1,107	1,226	1	0,268
width	0,291	0,190	1,528	2,336	1	0,126
color (2)	0,131	0,742	0,177	0,031	1	0,860
color (3)	-0,161	0,780	-0,206	0,043	1	0,836
color (4)	-1,245	0,855	-1,456	2,119	1	0,145

Cuadro: Model Summary - y

Model	Deviance	AIC	BIC
H ₀	225,759	227,759	230,912
H ₁	192,892	198,892	208,352

Cuadro: Coefficients

	Estimate	Standard Error	z	Wald Test		
				Wald Statistic	df	p
(Intercept)	-9,355	3,528	-2,652	7,031	1	0,008
width	0,307	0,182	1,686	2,843	1	0,092
weight	0,834	0,672	1,241	1,541	1	0,214

Cuadro: Model Summary - y

Model	Deviance	AIC	BIC
H ₀	225,759	227,759	230,912
H ₁	194,453	198,453	204,759

Cuadro: Coefficients

	Estimate	Standard Error	z	Wald Test		
				Wald Statistic	df	p
(Intercept)	-12,351	2,629	-4,698	22,075	1	< .001
width	0,497	0,102	4,887	23,887	1	< .001

Cuadro: Coefficients

	Estimate	Standard Error	Odds Ratio	Lower bound	Upper bound
(Intercept)	-12,351	2,629	$4,326 \times 10^{-6}$	0,000	0,001
width	0,497	0,102	1,644	1,347	2,007

Cuadro: Confusion matrix

Observed	Predicted		% Correct
	0	1	
0	28	34	45,161
1	16	95	85,586
Overall % Correct			71,098

Cuadro: Performance metrics

	Value
Accuracy	0,711
AUC	0,749
Sensitivity	0,856
Specificity	0,452
Precision	0,736

Curva ROC, logística simple

