Applied Biostatistics

Final project: GLM for predicting the number of satellites around pairs of horseshoe crabs

Lucia Montero Sanchis

1 Introduction

During the breeding season, male horseshoe crabs often crowd around a *nesting couple* or *pair* – a female with an attached male – and compete for fertilization [1]. It has been observed that some couples gather a large number of satellites around them, whereas others are ignored. This nonrandom distribution was tested by means of experimental manipulations, showing that the most likely reason behind the number of satellites was the characteristics of the mating females [2]. In experimental studies, females with a larger number of satellites were found to be larger and in better condition [2].

The dataset contains information for female crabs for which we want to predict the number of male satellites. We carry out an exploratory analysis and apply GLM to determine whether the characteristics of the female horseshoe crabs have a significant effect on the number of satellites.

2 Exploratory Data Analysis

This dataset contains information for 173 female horseshoe crabs. The explanatory variables are:

- **Weight (weight)**: Numerical variable representing the total weight of the crab in kilograms.
- **Carapace width (width)**: Numerical variable representing the carapace width in centimeters.
- **Color (color)**: Categorical, ordinal variable. There are four categories: 1 (light medium), 2 (medium), 3 (dark medium) and 4 (dark).
- **Spine condition (spine)**: Categorical, ordinal variable representing the condition of the spines. There are three categories: 1 (both good), 2 (one worn or broken) and 3 (both worn or broken).

In addition, we have the dependent variables:

- **Number of satellites (sat)**: Number of satellites observed around the female.
- Binary variable **(y)** representing whether the female crab had at least one satellite (1=yes, 0=no).

The analysis carried out focuses on the number of satellites (**sat**), whereas the binary variable representing the presence of satellites (**y**) will not be used. The variables color and spine condition are considered as factors throughout the entire analysis.

Table 1 shows the numerical summaries of the variables. We can see that the third quartile for variable **sat** is 5, meaning that at least 75% of crabs have 5 satellites or fewer. The histogram for the number of satellites (Figure 1) shows that over a third of crabs had 0 satellites. Satellite counts of 3 through 6 have slightly higher frequencies; for a larger number of satellites, the frequencies decrease considerably.

	sat	weight	width	color	spine
1	Min. : 0.000	Min. :1.200	Min. :21.0	1:12	1: 37
2	1st Qu.: 0.000	1st Qu.:2.000	1st Qu.:24.9	2:95	2: 15
3	Median : 2.000	Median :2.350	Median :26.1	3:44	3:121
4	Mean : 2.919	Mean :2.437	Mean :26.3	4:22	
5	3rd Qu.: 5.000	3rd Qu.:2.850	3rd Qu.:27.7		
6	Max. :15.000	Max. :5.200	Max. :33.5		

Table 1: Numerical summaries of variables.

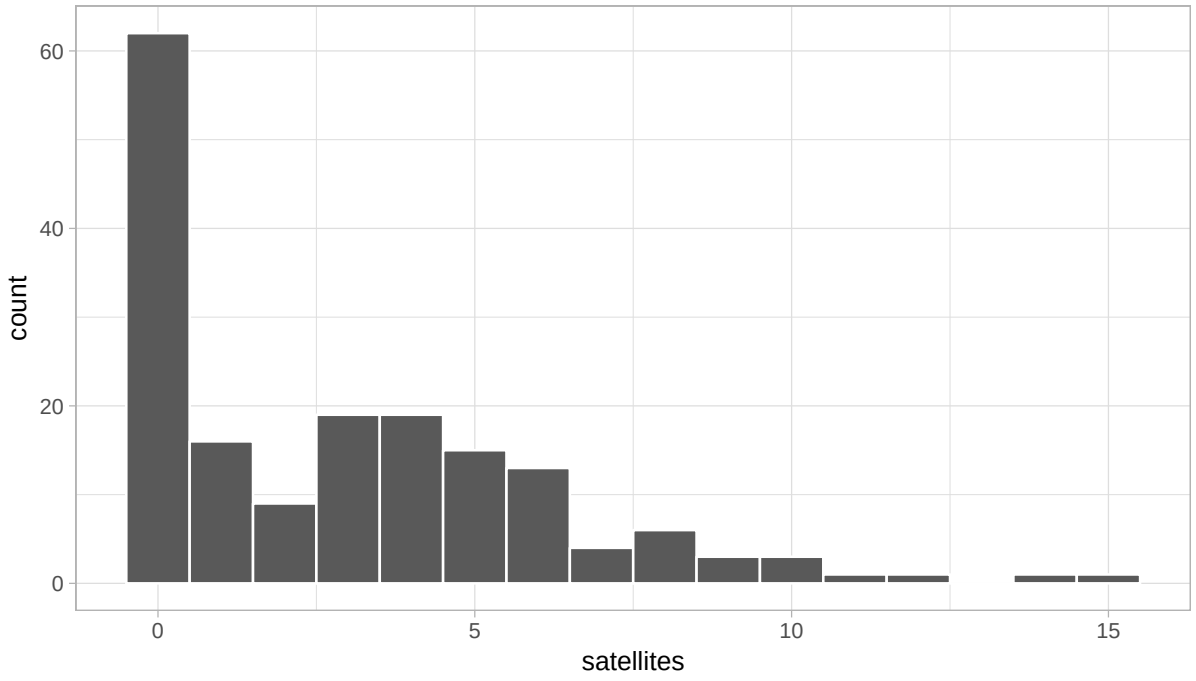
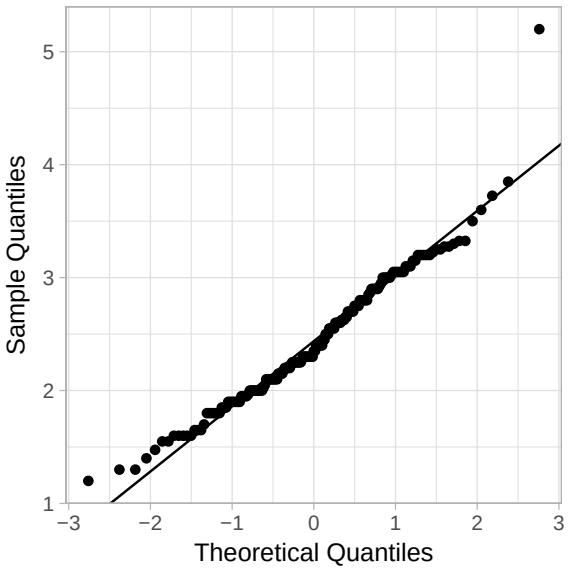


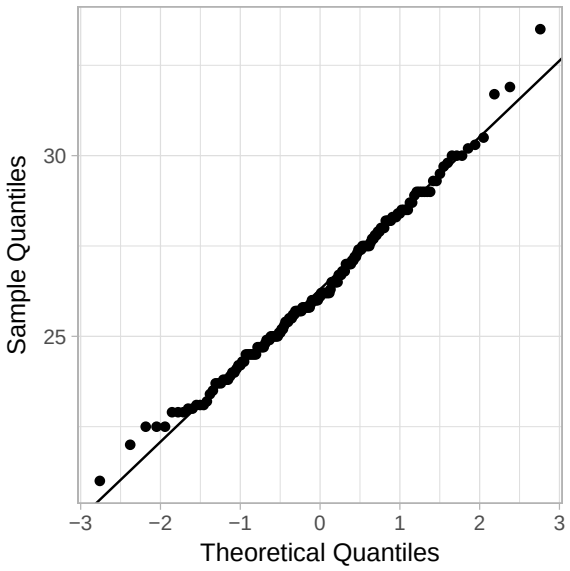
Figure 1: Histogram of the number of satellites.

We can use quantile-quantile (Q-Q) plots for determining whether **weight** and carapace **width** follow a normal distribution. Figures 2a and 2b show the normal Q-Q plots for weight and width, respectively. For both variables, the points follow a nearly linear pattern. Therefore, both variables seem approximately normally distributed.

There is however an outlier point for weight, that corresponds to a crab weighting 5.2 kg.



(a) Normal Q-Q plot for weight.



(b) Normal Q-Q plot for width.

Figure 2: Normal Q-Q plot for weight and width.

We check whether there is a linear association between the variables weight and width by computing Pearson's correlation coefficient, obtaining $r = 0.89$. This value suggests a strong positive linear relationship between the two variables. This relationship can be visualized in the scatterplot in Figure 3, which has been represented together with the linear model:

$$\text{width} = \beta_0 + \beta_1 \cdot \text{weight}$$

The summary of this linear model is shown in Table 2. For each term we test the null hypothesis that the coefficient is equal to zero. Conversely, the alternate hypothesis would be that the coefficient is different from zero:

$$H : \beta_0 = 0 \qquad A : \beta_0 \neq 0 \qquad ; \qquad H : \beta_1 = 0 \qquad A : \beta_1 \neq 0.$$

As seen in Table 2, both p-values are very small ($< 2e-16$). Therefore, we reject the null hypotheses that β_0 and β_1 are equal to 0, meaning that both terms are significant predictors of width. This is especially relevant in the case of weight, since there could be collinearity problems in the following section.

Regarding the values of the parameters, we can see in Table 2 that $\beta_1 = 3.24$. This value is consistent with the positive correlation coefficient that we has been presented, meaning that for each kilogram increment in weight we can expect an increase of 3.24 centimeters in carapace width.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.40	0.32	56.89	0.00
weight	3.24	0.13	25.10	0.00

Table 2: Results of fitting the linear model: width $\sim \beta_0 + \beta_1 \cdot$ weight

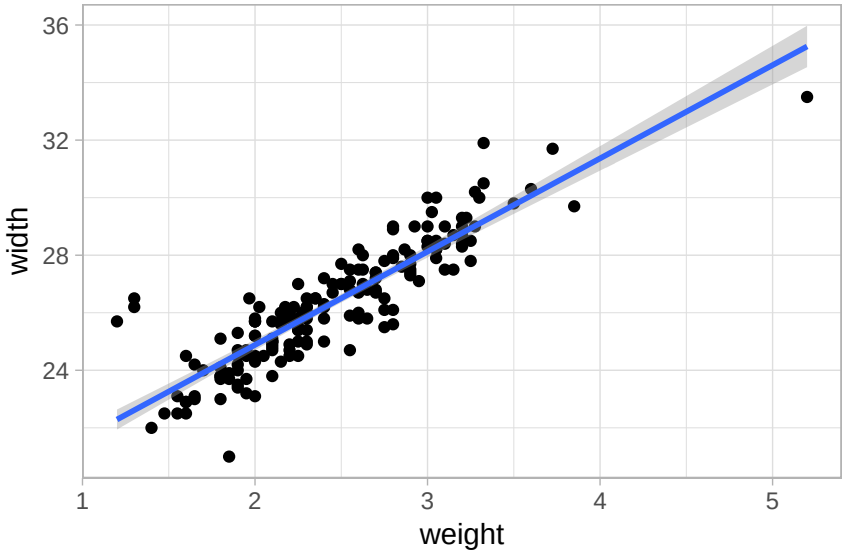


Figure 3: Scatterplot and regression line for width $\sim \beta_0 + \beta_1 \cdot$ weight. The shaded area represents the 95% confidence region. The model is described in Table 2.

Table 1 shows that 95 out of 173 crabs have **color** 2, representing 55% of the total number of crabs. We can also see that the great majority of crabs have both spines worn or broken, with 70% of crabs having value 3 for variable **spine**.

Figure 4 gives an overview of the fraction of crabs with each color and carapace width combination. The proportion of crabs with the darkest colors (3 and 4) increases with the number of worn or broken spines, whereas the proportion of crabs with the lightest color (1) decreases. The proportion of crabs with a medium coloring (2) remains nearly the same for the three possible spine conditions. For instance, we can see that there are very few light-colored crabs in the group with both spines worn or broken. Shell coloring depends on age, with older horseshoe crabs being darker [3]. This could explain why these results are observed, since older crabs could have had more chances of breaking or wearing down their spines.

3 Models

The number of satellites is a count, which can only take on positive values and is unlikely to be normally distributed. Since multiple regression requires normally distributed variables, this would not be a reasonable approach. Instead, we can fit a **Generalized**

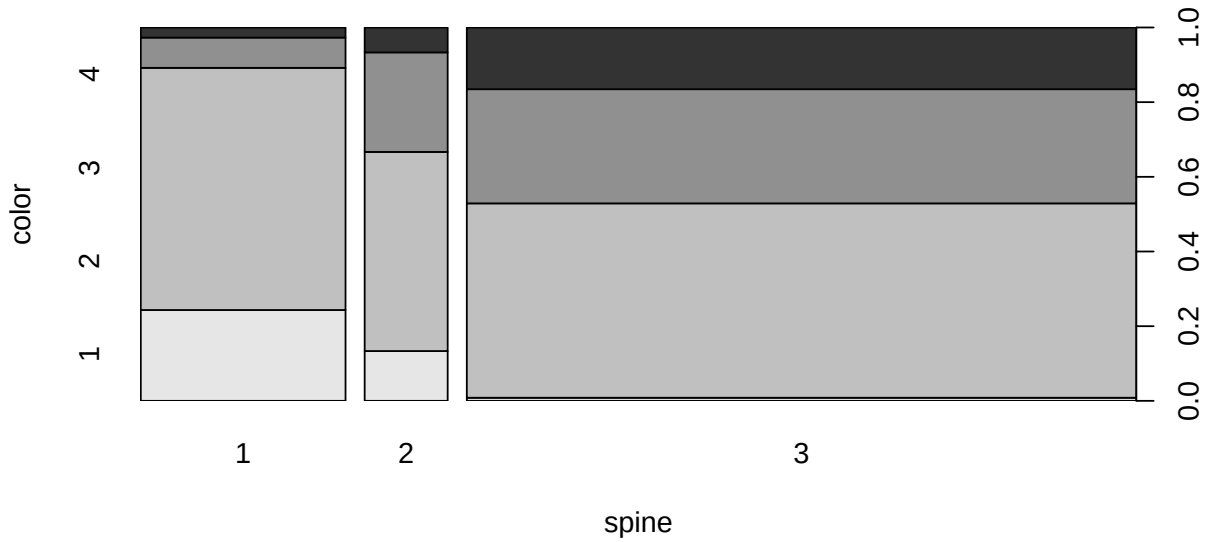


Figure 4: Mosaic plot displaying carapace color and spine condition. The colors used correspond with the color group (1 being the lightest, 4 the darkest).

Linear Model (GLM) with a **log link function** and a **Poisson error distribution**; this is known as **Poisson regression**. A GLM is a more flexible generalization of the ordinary linear regression, that allows for response variables with errors not normally distributed. This is desirable because, since we are predicting a count, it is not likely for our response variable to be normally distributed. However, Poisson regression assumes that the response variable has a Poisson distribution. If this assumption is not true for the data, then other approaches should be considered. One of the characteristics of the Poisson distribution is that the mean is equal to the variance; if there is over-dispersion, meaning that the variance is greater than the mean, a different approach should be taken. It is also possible that the Poisson regression does not fit well if there is an excess number of zeros in the count data.

Since weight and carapace width have a correlation of 0.89, only weight will be used to avoid collinearity problems. They are strongly correlated, so this decision should not affect strongly the quality of the model fit. We will go back to this decision to verify which of the two variables should be used.

3.1 Poisson regression

One of the distributions used for modeling count data is the Poisson distribution. The results of fitting a **Generalized Linear Model** (GLM) with a **log link function** and a **Poisson error distribution** are shown in Table 3. Using a log link function $g(\mu) = \log(\mu)$ ensures that the fitted values are positive [4].

Although some of the predictors are significant, the model from Table 3 does not seem to

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.04	0.25	-0.17	0.87
weight	0.55	0.07	7.48	0.00
color2	-0.27	0.17	-1.60	0.11
color3	-0.52	0.19	-2.68	0.01
color4	-0.54	0.23	-2.40	0.02
spine2	-0.16	0.21	-0.76	0.45
spine3	0.09	0.12	0.76	0.45

Table 3: Results of fitting a GLM with a log link function and with a Poisson error distribution to explain the number of satellites.

fit well because residual deviance is much larger than the number of degrees of freedom:

Residual deviance: 549.70 on 166 degrees of freedom

The ratio between residual deviance and degrees of freedom is $549.70/166 = 3.31$, which is much larger than 1. In Section 3.2 we consider different alternatives to obtain a better fit.

3.2 Dealing with over-dispersion and excess number of zeros

There are different approaches for dealing with the limitations of Poisson regression. In the case of over-dispersion, we can use a **quasipoisson** model. This approach differs from the previous in leaving the dispersion parameter unrestricted; it assumes $Var(Y_i) = \phi\mu_i$ and estimates ϕ from the data [4]. Another way is to use **Negative Binomial** regression [4].

Table 4 presents the results obtained when fitting a **quasipoisson model** to the number of satellites. Compared to Poisson regression, the dispersion parameter (scale parameter) is no longer assumed to be $\phi = 1$; the dispersion parameter for quasipoisson is estimated from the data, obtaining $\phi = 3.21$. After testing the null hypothesis that the coefficient equals 0 ($H : \beta = 0$, $A : \beta \neq 0$), in this case only weight is significant at level 0.05.

When looking at the residual deviance and number of degrees of freedom for this model, we see that they are the same as for the Poisson model. Therefore, the quasipoisson model does not fit well either.

Residual deviance: 549.70 on 166 degrees of freedom

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.04	0.45	-0.09	0.93
weight	0.55	0.13	4.17	0.00
color2	-0.27	0.30	-0.89	0.37
color3	-0.52	0.35	-1.50	0.14
color4	-0.54	0.40	-1.34	0.18
spine2	-0.16	0.38	-0.42	0.67
spine3	0.09	0.21	0.42	0.67

Table 4: Results of fitting a quasipoisson model to explain the number of satellites.

It is also possible to model count data assuming a **Negative Binomial** distribution. Compared to the quasipoisson approach, one of the advantages of using Negative Binomial regression is that the Akaike Information Criteria (AIC) can be used to obtain a reduced model. This criterion estimates the relative quality of statistical models for a certain dataset. The results of fitting a Negative Binomial regression model are shown in Table 5. The residual deviance and number of degrees of freedom obtained for the Negative Binomial regression are:

Residual deviance: 196.51 on 165 degrees of freedom

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.32	0.56	-0.57	0.57
weight	0.69	0.17	4.18	0.00
color2	-0.32	0.37	-0.86	0.39
color3	-0.60	0.42	-1.43	0.15
color4	-0.58	0.46	-1.25	0.21
spine2	-0.24	0.39	-0.61	0.54
spine3	0.04	0.25	0.17	0.86

Table 5: Results of fitting a Negative Binomial regression model to explain the number of satellites.

The ratio between residual deviance and degrees of freedom is $196.51/165 = 1.19$, which gives a better fit than both the Poisson and quasipoisson models. It is also worth mentioning that the dispersion parameter estimated for Negative Binomial (θ) is 0.97; it is taken to be 1 in the model. In this case, when testing the null hypothesis that the coefficient is equal to zero (with the alternate hypothesis that the coefficient is different from zero, $H : \beta = 0$, $A : \beta \neq 0$), variable weight is the only predictor with a p-value lower

than 0.05. Therefore, we reject the null hypothesis that the coefficient for weight equals 0, meaning that weight is a significant predictor of the number of satellites at 0.05 level. The variable weight has a coefficient of 0.69, so that for each kilogram increase in weight, the expected log count of the number of satellites increases by 0.69.

The residual deviances and numbers of degrees of freedom for the three models are summarized in Table 6.

	Rdev	df	Rdev/df
Poisson model	549.59	165	3.33
Quasipoisson model	549.59	165	3.33
Negative Binomial model	196.51	165	1.19

Table 6: Residual deviance (Rdev), degrees of freedom (df) and residual deviance over degrees of freedom ratio (Rdev/df) for the considered models.

3.3 Dealing with collinearity

In the beginning of the section we decided to use all the predictors except width, because of the strong correlation between width and weight. However, either width or weight could have been taken out of the equation. We now use AIC to compare the Negative Binomial (NB) models with each one taken out, both as models reduced by backwards stepwise elimination and full models [5]. The values are summarized in Table 7.

	Without **weight**	Without **width**
Full models	764.33	761.32
Reduced models	757.29	**754.64**

Table 7: AIC values for full NB models and NB models reduced by backwards stepwise elimination, after taking out either weight or width variables. Bold value is the lowest AIC.

We are interested in the model with the lowest AIC, and we can see from Table 7 that the lowest AIC is achieved by removing width from the model. This is the case for full models and for models reduced by backwards stepwise elimination. This consists on starting with all variables and testing the deletion of each variable using a chosen criterion. The variable whose loss gives the most statistically insignificant deterioration of the model fit is deleted. This process is repeated until there are no variables that can be deleted without a statistically significant loss of fit.

Table 8 shows the results of fitting the model reduced by backwards stepwise elimination after having removed variable width, obtaining the model:

$$\text{sat} = \exp\left(\beta_0 + \beta_1 \cdot \text{weight}\right), \text{ with } \beta_0 = -0.86 \text{ and } \beta_1 = 0.76$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.86	0.40	-2.14	0.03
weight	0.76	0.16	4.82	0.00

Table 8: Results of fitting a Negative Binomial regression model to explain the number of satellites and then applying backwards stepwise elimination.

When testing the null hypothesis that a coefficient is equal to zero (with the alternate hypothesis that the coefficients are different from zero), we find that both coefficients have a p-value lower than 0.05. This means that weight is a significant predictor of the number of satellites at a 0.05 significance level. The coefficient from the reduced model for variable weight equals $\beta_1 = 0.76$; hence the expected log count of the number of satellites increases by 0.76 for each kilogram increase in weight. It is also worth mentioning that the dispersion parameter is again taken to be 1, and the estimated value in this case is $\theta = 0.93$.

In order to visualize the fitted model, it is shown in Figure 5 together with the scatterplot of the data.

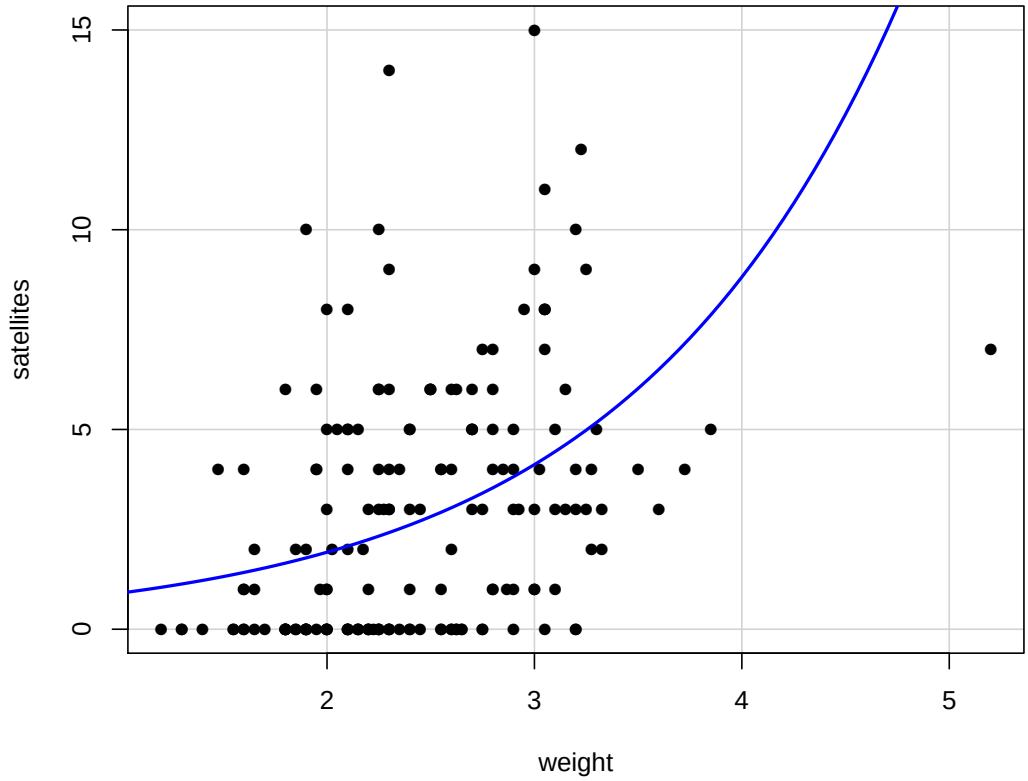


Figure 5: Scatterplot and Negative Binomial regression model for predicting the number of satellites. The model is described in Table 8.

It does seem to fit the data, although the scatterplot shows that the points are quite sparse.

4 Conclusions

Of the three models that have been used to fit the data, Negative Binomial regression with all the variables was the one with the best fit. We then reduce the full model by applying backwards stepwise elimination using AIC and obtain a reduced model with weight as the only predictor.

Based on our analysis, we conclude that heavier female crabs tend to gather a larger number of satellites. Given the strong positive correlation found between carapace width and weight, this implies that wider crabs tend to be heavier and therefore have more satellites.

From the analyses carried out, color and spine condition do not seem to affect the number of satellites.

Using Negative Binomial regression we obtained a model that seemed to fit the data. However, the points were quite sparse.

References

[1] H. J. Brockmann, C. Nguyen, and W. Potts, "Paternity in horseshoe crabs when spawning in multiple-male groups," *Animal Behaviour*, vol. 60, no. 6, pp. 837–849, 2000.

[2] H. J. Brockmann, "Satellite male groups in horseshoe crabs, Limulus polyphemus," *Ethology*, vol. 102, no. 1, pp. 1–21, 1996.

[3] H. Cook and C. E. Matthews, "Lessons from a "living fossil"," *Science and Children*, vol. 36, no. 3, p. 16, 1998.

[4] A. Zeileis, C. Kleiber, and S. Jackman, "Regression models for count data in R," *Journal of statistical software*, vol. 27, no. 8, pp. 1–25, 2008.

[5] D. Rossiter, "Tutorial: An example of statistical data analysis using the R environment for statistical computing," 2008.