A Bayesian Analysis of Factors Influencing Number of Satellite Males in Horseshoe Crabs During Breeding Season

Statistical Methods

Horseshoe crabs are marine and brackish water arthropods which live primarily in and around shallow coastal waters on soft, sandy or muddy bottoms. Despite their name, horseshoe crabs are more closely related to arachnids and are not true crabs or crustaceans. During breeding season, horseshoe crabs arrive on the beach in pairs to spawn. Additional unattached males called "satellite males" crowd around the nesting couples and compete with attached males for fertilisations. These satellite males form large groups around some couples while ignoring others. In this analysis we look at characteristics of the nesting female crabs as potential factors explaining the number of satellite males surrounding a nesting couple.

The dataset contains 173 observations on 5 variables. There are 2 categorical variables, color (indicating the color of the female) and spine (indicating the condition of the spines of the female), 2 continuous numeric variables, weight (weight in kgs of female) and width (carapace width in cms), and the response variable satellites (number of satellites) which is a non-negative integer.

To address the number of zeros in the dataset, a number of different models were assessed including Possion and Negative Binomial models while applying hurdles and zero inflation techniques. When observing the diagnostic plots there also seemed to be some curvature in the data so polynomials were introduced into the model formula to account for this. Models were assessed using diagnostic plots, accuracy in posterior prediction plots, and comparing models using ELDP. The final model was a zero-inflated negative binomial model with formula

$$\text{satellites} \quad \text{weight}^3 + \text{color} + \text{spine} + \text{width}$$

Results

The model was fit using the brm() function from the brms package with the family argument set to zero_inflated_negbinomial and using uninformative priors (N(0,10)) for the parameters of the model. The model produces the following summary information:

```
##  Family: zero_inflated_negbinomial
##   Links: mu = log; shape = identity; zi = identity
## Formula: satellites ~ poly(weight, 3) + color + spine + width
##    Data: CrabSatellites (Number of observations: 173)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept        4.57      2.17     0.32     8.83 1.00     2078     2606
## polyweight31     6.84      2.20     2.57    11.15 1.00     2089     2623
## polyweight32    -2.76      0.96    -4.68    -0.91 1.00     2924     2595
## polyweight33     2.09      0.98     0.22     4.08 1.00     3445     2820
```

```
## colormedium        -0.09    0.27   -0.63    0.45 1.00    1890    2359
## colordarkmedium     -0.35    0.30   -0.94    0.23 1.00    1896    2352
## colordark            0.43    0.39   -0.36    1.18 1.00    1979    2248
## spineonebroken      -0.24    0.31   -0.86    0.37 1.00    3946    3050
## spinebothbroken     -0.14    0.18   -0.50    0.23 1.00    2755    2947
## width               -0.12    0.08   -0.28    0.04 1.00    2031    2661
##
## Further Distributional Parameters:
##        Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     5.71      3.35     2.29    13.15 1.00     2553     2181
## zi        0.30      0.04     0.21     0.38 1.00     3254     2124
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The above output shows the parameter estimates for our bayesian regression model. The intercept shows the baseline estimated log count with both spines in good condition, light medium color and mean values of weight and width. This equates to 90 satellites which doesn't seem consistent with the data collected. The high estimated error and wide 95% CI for the intercept parameter lowers my confidence in the accuracy of the model for prediction. The different parameters for the different levels of spine and color show the estimated change in the log counts of satellites when changing from the reference level, for example, there is an increase of 0.42 in the estimated log counts for a female crab with dark color compared to a female crab with light medium color. The width parameter estimate gives the estimated change in log counts of satellites resulting from a one unit increase in carapace width (cm). The effects are visualised in the following plots.

```
## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail

## Warning in glmmTMB::glmmTMB(formula = cond_formula, ziformula = ziformula, :
## some components missing from 'family': downstream methods may fail
```
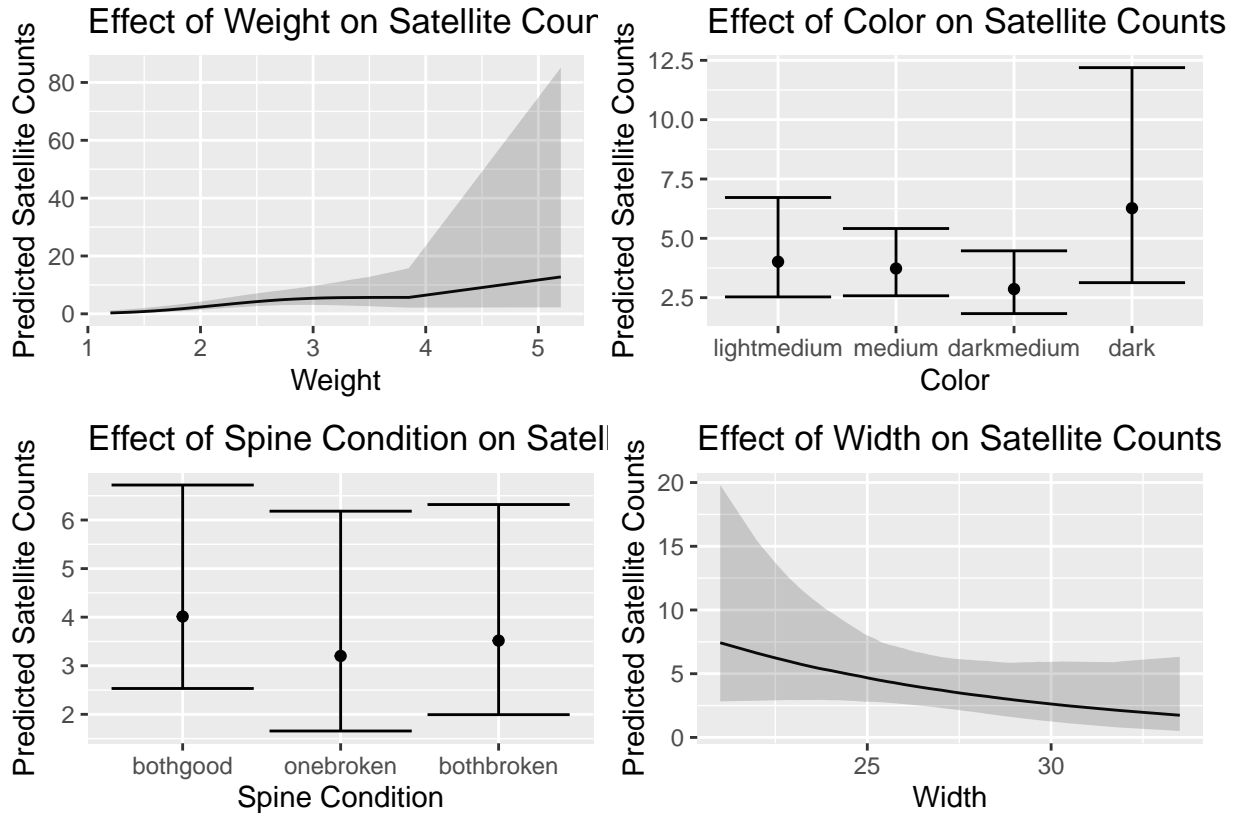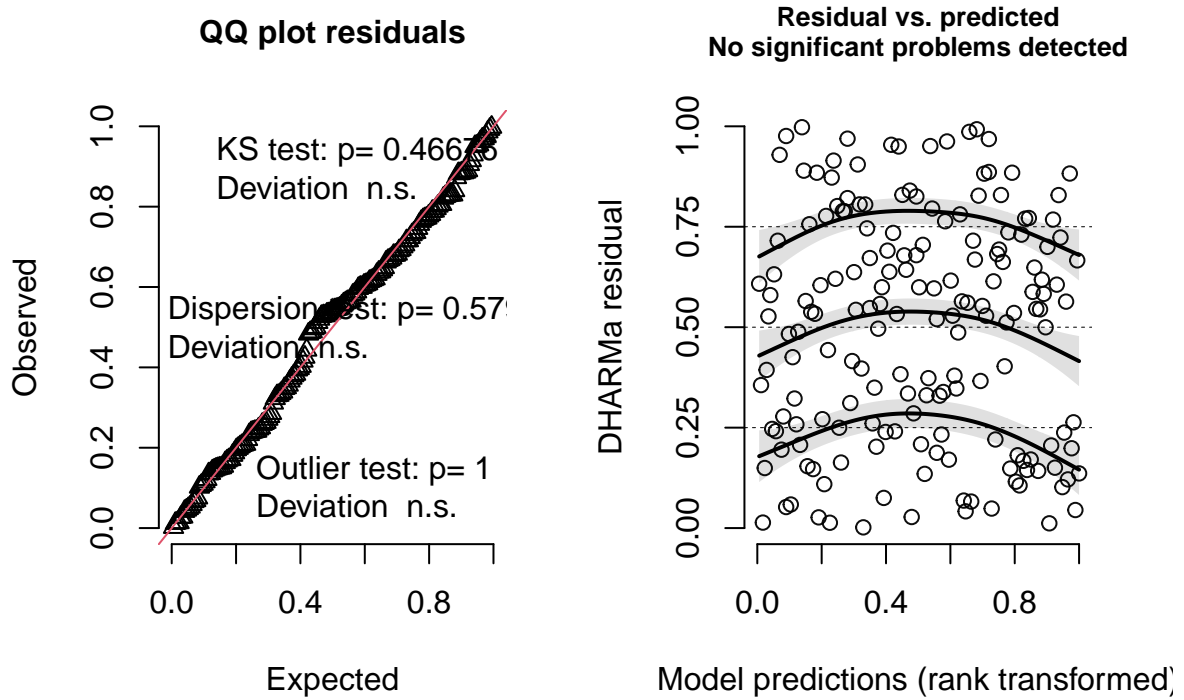
The model shows that there is a positive relationship between female crab weight and the number of satellite males however the 95% CI is quite large, especially for females at the larger end of the weight scale. This could be due to the small sample size of large females in the dataset. The width of the carapace of the female crab seems to have a negative relationship with number of satellites with the wider the carapace, the less satellite males attracted. For the effects of the categorical variables, both spines being good seems to attract the most satellites however, the change between the levels of spine condition is not great. The color of the crab shows a greater discrepancy with dark colored females attracting greater numbers of satellites than the other 3 colors.

When assessing model assumptions and fit, the DHARMa pakage was used. The following plot shows the results from this.

DHARMa residual

QQ plot residuals

KS test: p= 0.4667
Deviation n.s.

Dispersion test: p= 0.57
Deviation n.s.

Outlier test: p= 1
Deviation n.s.

Observed

Expected

Residual vs. predicted
No significant problems detected

DHARMa residual

Model predictions (rank transformed)

As you can see above, the QQ residual plot follows the line quite well suggesting no problems with the residuals. The residuals vs predicted plot shows some curvature in the data even after adjusting by using polynomials in the model formula although no significant issues are present. I also checked posterior prediction fit with the results in the following plot.
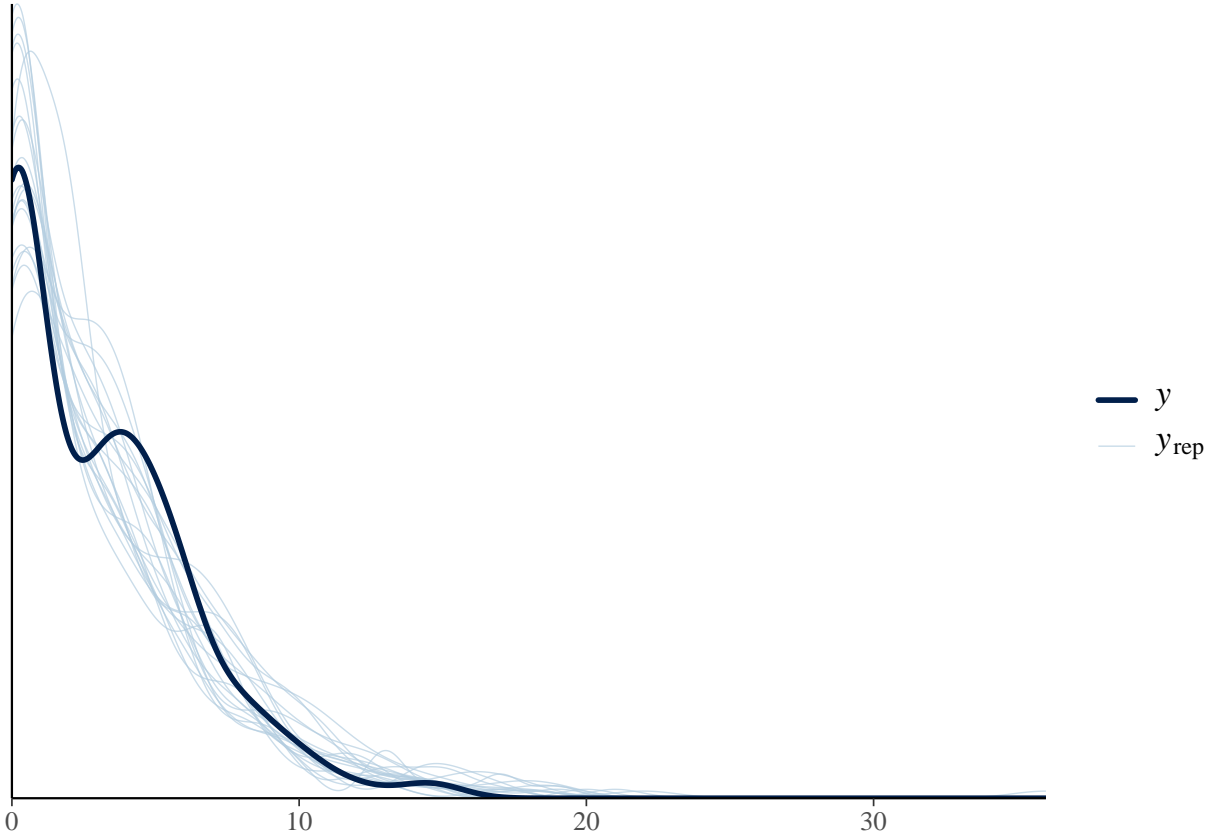
The posterior draws seem to generally follow our model quite well, although there is certainly some room for improvement.

Problems encountered

The main problems encounter with this analysis were how to the problem of excess zeros in the data and how to deal with the curvature in the data. For the excess zeros, as mentioned above, I tried numerous models using both poisson and negative binomial distributions for the response variable and using either zero hurdle models or zero inflated models before settling on the final model. I also introduced quadratic and cubic polynomials with the continuous explanatory variables within the model to address the curvature in the data before settling on the final formula. Furthermore, I changed the structure of the spine and color variables from ordinal factors to non-ordinal factors to help in the interpretation of the results.