Exploration 6.2: Horseshoe crabs

Learning Goals:

- Explain the motivation and need for logistic regression
- Utilize a logistic regression model using categorical or quantitative explanatory variables

Background: Horseshoe crabs are arthropods that live in salty water. Researcher Brockman (1996) studied the characteristics of female horseshoe crabs and how they were associated with whether and how many male horseshoe crabs were attached (called "satellites") to the female.

The dataset **Horseshoecrabs.txt** (retrieved from https://users.stat.ufl.edu/~aa/cda/data.html on May 26, 2022) contains data on variables such as: whether any satellites are attached, how many satellites are attached, width of the shell (cm), and color of shell. In this dataset, color is recorded as dark or medium. These data will be used to investigate whether the color of the shell and the width of the can predict whether the female will have a satellite.

STEP 1: Ask a research question.

1. What appears to be the primary response variable in this study? Is the response variable quantitative or categorical?

2. Identify three explanatory variables of primary interest. Classify each as quantitative or categorical.

3. Complete the Sources of Variation diagram below, including conjecturing at least one additional source of (unexplained) variation. Was this an observational study or an experiment?

Observed Variation in:	Sources of explained variation	Sources of unexplained variation
	• •	•
Design *Inclusion criterion*		

STEP 2: Design a study and collect data.

Open the **Horseshoecrabs** datafile.

4. Do the data arise from an observational study or a randomized experiment? How are you deciding?

5. What is the sample size in the dataset?

STEP 3: Explore the data.

Use statistical software to make a graph of the *have.satellite* response variable.

Note that the *have.satellite* variable indicates whether or not a female horseshoe crab has at least one satellite.

6. How many females have satellites? What is the proportion who have satellites?

7. Carry out a descriptive analysis (graphs, numerical summaries) exploring the association between *have.satellite* and *color* (you should include and report an odds ratio). What theory-based test could we use to test the statistical significance of this association? Do you think the p-value will be large or small? Explain.

8. Carry out a descriptive analysis (graphs, numerical summaries) exploring the association between *have.satellite* and *width*. What theory-based test could we use to test the statistical significance of this association? Do you think the p-value will be large or small? Explain.

9. Use statistical software to carry out the tests suggested in the previous two questions. (Include a confidence interval for the population odds ratio for *color*.) Does either explanatory variable explain significant variation in whether or not a female horseshoe crab has a satellite?

10. Suggest one or two limitations to these analyses.

One limitation to the above analyses is that they consider each variable in isolation. Because this is an observational study, there could be covariation between *color* and *width*. Even if there wasn't, the association between presence of satellites and width of the shell could look different for females with darker shells and females with lighter shells. We would like to develop a *model* that allows us to predict presence of satellites based on both variables together. First, we consider modelling a *trend* for the proportion that have at least one satellite at different shell widths.

Use software to compute the proportion that have at least one satellite for each width value. Then produce a graph of the "yes" percentages vs. the width values.

11. Describe the association. Would you consider it linear? Are there any difficulties with using a line to model the relationship between the proportion have at least one satellite and width? (*Hint:* What happens for widths greater than 34cm?) Explain.

Key Idea: When working with proportions, you may be able to fit a linear model, but often a linear model is not the most appropriate. The expected S-shaped curve (as probabilities must be bounded by 0 and 1) implies a power transformation (e.g., x^2, log(x)) and is not likely to be helpful either. Instead, we can try a ***logit transformation***: $\ln(\pi/(1-\pi))$, where we use π to represent the population proportion of successes. This is equivalent to using *log odds* as the response variable.

Many statistical packages will perform this transformation for you when you specify a binary variable as your response variable. A method similar to "least squares estimation" (actually termed "maximum likelihood estimation") then estimates the appropriate intercept and slope coefficients for this ***logistic regression model***.

Definition: *Logistic regression* models use quantitative or categorical explanatory variables to predict the probability of success for a binary response variable by fitting the *expected log odds of success* as $\beta_0 + \beta_1 x$.

Suppose we run a logistic regression model predicting presence of satellites by age and find *predicted log odds of having at least one satellite* = -12.351 + 0.497 × *width*. Notice that this corresponds to *predicted odds of having a satellite* = $e^{-12.351 + 0.497 width}$ = $(0.000004)(1.644^{width})$.

12. Using this equation, what are the predicted odds of having a satellite when *width* = 0? How could we change the *width* variable so that the intercept is more meaningful in our model?

13. What are the predicted odds of having a satellite for a female horseshoe crab with a shell that is 25 cm wide? Use the predicted odds to predict the probability that a female horseshoe crab with a shell width of 25 cm has at least one satellite. (*Hint*: You will need to rearrange the equation.)

14. To help interpret the slope in the equation (0.497), set up and simplify the ratio of $(0.000004)(1.644^{width+1})$to $(0.000004)(1.644^{width})$. How does the predicted odds change when shell width increases by one cm?

Key Idea: The slope of a logistic regression model indicates the multiplicative change (× e^{slope}) with a one-unit increase in the explanatory variable.

15. How much would the odds ratio change if we compared females whose shell widths differed by 2 cm.?

What about the categorical explanatory variable?

Suppose we code the *color* variable as 1 for dark and 0 for medium, and we run a logistic regression model to predict the log-odds of having a satellite from only this indicator variable. We find

predicted log odds of having a satellite = 0.989 – 0.989 × *color(dark)*

16. What is the predicted probability of having a satellite for crabs with medium shells? (*Hints*: What is the value of *color* for dark shells? First find the predicted log odds, then the predicted odds, then the predicted probability.)

17. What is the predicted odds ratio of having a satellite for female horseshoe crabs with medium shells compared to females with dark shells? (*Hint*: What do we mean by a one-unit change for this indicator variable?) Which color shell is more likely to have a satellite? How does this correspond to the sign of the color slope coefficient?

18. Verify that the estimate of the odds ratio you found in the previous question matches the calculation of the sample odds ratio from the two-way table earlier.

STEP 4: Draw inferences beyond the data.

A key advantage to the logistic model is automatic generation of p-values and confidence intervals.

19. Use statistical software to fit two logistic regression models: one with *color* as the explanatory variable and one with *width* as the explanatory variable. Fill in the Regression tables below with your results.

Fit a logistic regression model to predict *have.satelllite* from *color*:

Regression table for *width (quantitative)*

Term	Coeff	SE	Chi-square	*p*-value
Intercept				
Width				

Regression table for *color (did you use indicator or effect coding?)*

Term	Coeff	SE	Chi-square	*p*-value
Intercept				
Color				

One interesting observation is these are no longer *t*-tests but actually chi-square tests. (You can verify the chi-square values, apart from rounding, by finding *z* = coeff/SE and then squaring.) However, that detail isn't important and you will evaluate the p-values as with any other regression table.

20. Does a female crab's shell width explain a significant amount of variation in whether or not the female has a satellite? Does a female crab's shell color explain a significant amount of variation in whether or not the female has a satellite? Clearly explain how you are deciding.

21. Which variable provides stronger evidence of an association with whether or not a female horseshoe crab has a satellite? How are you deciding?

22. Use the logistic regression model to produce a 95% confidence interval for the slope coefficient of *color*. (*Hint*: If your technology does not do this automatically, think about how you can approximate a 95% confidence interval using the slope coefficient and its standard error).

Key Idea: Given the confidence interval for the slope coefficient in a logistic regression model, exponentiating the endpoints gives you a confidence interval for the population odds ratio.

23. Use the logistic regression model to produce and interpret a 95% confidence interval for the population odds ratio for *color*.

24. Compare your analysis of the association between *color* and *have.satelllite* to chi-square tests and confidence intervals for the odds ratio. Include the new output and summarize what you learn.

Now we can just as easily determine a confidence interval for the odds multiplier associated with width, keeping *have.satellite* as the response variable.

25. From the software output, determine a confidence interval for the slope of *width*. Produce and interpret a 95% confidence interval for the odds multiplier for *width*.

STEP 5: Formulate conclusions.

26. Summarize what you have learned from your analysis so far. Your discussion should address significance, estimation (confidence intervals), causation, and generalizability.

STEP 6: Look back and ahead.

27. Identify at least one concern you have with this analysis. Explain why it is a concern and what could be done to address your concern.