Exploration 5.3: Predicting Skeletal Mass from Length of Humerus in Birds

LEARNING GOALS

- Fit polynomial model
- Assess when a polynomial model is appropriate

Background: Can an organism's skeletal mass be predicted by various features of the skeletal structure? Can we predict how large a bird might have been if we find say a part of the skeleton? Martin-Silverstone et al. (2015) assembled data on several species of birds to examine the association between total body mass and skeletal mass. We will look at a subset of 73 male birds to explore the relationship between skeletal mass (in kilograms) and length of the humerus (in millimeters). The humerus is the upper "armbone." In particular, we want to predict skeletal mass from humerus length.

The dataset ***SkeletalDataMales*** contains information on the skeletal mass and humerus length of for 73 male bird from various species.

1. Create a scatterplot of the skeletal mass (*skelmass*) vs. the humerus length (*humlen*). Describe any patterns you see. Are there any unusual observations? Does it make sense to fit a linear model to these data?

Definition: ***Polynomial models*** include terms x, x^2, x^3, … to help model nonlinear associations between two quantitative variables. The highest power used is the *degree* of the polynomial model. It is rare that that model will go beyond degree 2 (quadratic) or 3 (cubic). A polynomial model is still considered a "linear model" because it will be of the form $y = a + bx_1 + cx_2$ where $x_1 = x$ and $x_2 = x^2$ and so on… In fact, you can think of a quadratic term x^2 as x interacting with itself.

In a spreadsheet, create a new variable: humlen^2 (multiply the humlen column by the humlen column).

2. Fit a ***quadratic model*** to predict skeletal mass from *humlen* and *humlen*^{2}. Include a residual analysis for this model—does this model appear to adequately capture the curvature in the data? Provide justification for your answer.

3. Provide an interpretation of the intercept of this model. Does the intercept make sense in this context? Explain.

4. Write out the prediction equation for the quadratic model. Is the coefficient of *humlen*^{2} positive or negative, and what does that imply about the relationship between skeletal masss and humerus length as humerus length increases? (*Hint*: You can think of this as an interaction between *humlen* and *humlen*!)

We expect *humlen*^{2} and *humlen* to be related to each other, but for the range of data values at which we are looking, is the relationship linear?

5. Produce a scatterplot of *humlen*^{2} and *humlen.* What do you learn? How can we address this issue?

6. Use your model to find a 95% prediction interval on the predicted skeletal mass (in kg) of a bird that has a humerus length of 175 millimeters. Recall that a 95% prediction interval is computed as roughly $\hat{y} \pm 2$(*SE of residuals*). Interpret your interval in context.

> **Key Idea:** *Standardizing* the variable before applying a polynomial model moves the "curved" part of the association between the linear and quadratic terms into the middle of our explanatory variable region, reducing the linear association between the linear and quadratic terms.

Standardize the *humlen* variable. Then recreate the quadratic term using the standardized variable. Then include both *standardized humlen* and (*standardized humlen*)^2 in the model.

7. Examine a scatterplot of *standardized humlen* and (*standardized humlen*)^2. Is the association still linear?

8. Provide an interpretation of the intercept of this model (include units!). Does this intercept make sense in this context?

9. To help interpret this quadratic behavior, write out the slope of *std.humlen*^{2} in terms of *std.humlen*:

Slope of *std.humlen*^{2} = ____________ + ___________ *std.humlen*

a) Using the above expression, what is the slope of *std.humlen*^{2} when *std.humlen* = -2?

b) Using the above expression, what is the slope of *std.humlen*^{2} when *std.humlen* = 2?

c) How do these values compare? How is this related to the quadratic behavior we are modelling?

d) What is the predicted slope when *std.humlen* = 4? Why is this prediction problematic in this context?

10. Answer each of the following questions. Make sure you not only answer the question, but also make it clear how you are deciding

- Is the standardized quadratic model useful in explaining variation in skeletal mass?
- Is the standardized quadratic model statistically significant?
- Is the standardized quadratic model valid?
- What does the standardized quadratic model look like (describe the behavior/form of association)? Does this model make sense in context?
- What is your prediction for the skeletal mass of a bird with a humerus length of 175 mm using the standardized model? Do you find this prediction useful?

11. Create the cubic term: *std.humlen*^{2} × *std.humlen* and add this variable to the model. Does the cubic model explain significantly more variation in the skeletal mass than the quadratic model? How are you deciding?

12. Examine residual plots for the quadratic model as well as for the cubic model. Which model do you recommend if the goal is to predict the skeletal mass of birds from the length of their humerus? Does the model you are recommending appear valid? Comment on each validity condition and if you don't think the condition is met, comment on the possible consequences in interpreting the model.

13. Would the model you have created be *useful* in predicting the skeletal mass of birds from the length of their humerus? Explain why or why not. (*Hint*: Consider issues beyond validity conditions.)