Exploration 10.3: Predicting Total Body Mass from Skeletal Mass in Birds

Least Squares Regression

LEARNING GOALS

- Understand that one way a scatterplot can be summarized is by fitting the best-fit (least squares regression) line and interpreting both the slope and intercept of the best-fit line in the context of the two variables on the scatterplot.
- Find the predicted value of the response variable for a given value of the explanatory variable.
- Understand that slope = 0 means no association, slope < 0 means negative association, and slope > 0 means positive association. Further, the sign of the slope will be the same as the sign of the correlation coefficient.
- Understand that extrapolation is using a regression line to predict values outside of the range of observed values for the explanatory variable, including the special case of $y = 0$ when applicable.
- Understand the concept of residual and find and interpret the residual for an observational unit, given the raw data and the equation of the best-fit (regression) line.
- Understand the relationship between residuals and strength of association and that the best-fit (regression) line minimizes the sum of the squared residuals.
- Find and interpret the coefficient of determination (R^2) as the squared correlation and as the proportion age of total variation in the response variable that is accounted for by changes (variation) in the explanatory variable.
- Understand that influential points can substantially change the equation of the best-fit line and that observations with extreme values of the explanatory variable may potentially be influential.

An organism's body mass can help predict other variables like metabolism, growth rate, diet, locomotion model, etc. In particular, in birds, total body mass helps predict their type of flight. But how do we determine body mass of extinct birds, where all that we have access to is fossil remains? Martin-Silverstone et al. (2015) collected data on 487 birds to examine the association between total body mass and skeletal mass. We will look at a subset ($n = 30$) of this dataset to explore the relationship between total body mass (in grams) and skeletal mass (in grams). In particular, we want to predict total body mass from skeletal mass.

1. Let's think about this study.

    a. What are the observational units in this study?

    b. Identify the two variables of primary interest recorded for each observational unit. Which is the explanatory variable, and which is the response? Also classify these variables as categorical or quantitative.

    c. Is this an observational study or a randomized experiment?

Open up the **Corr/Regression** applet, clear out the results, and paste the **BirdMass** data into the data window. Click on **Use Data** and look at the scatterplot displayed. Verify that the scatterplot has the explanatory and response variables in the correct position. If they are not, click on the **(Response, Explanatory)** button to switch them around. Make sure to delete the last empty row if there is one.

2. Describe the association between the variables as revealed in the scatterplot. (*Hint*: Remember to comment on direction, strength, and form of the association as well as unusual observations.)

3. Would you say that a straight line could summarize the relationship between skeletal mass and total mass reasonably well?

4. Check the **Show Movable Line** box to add a blue line to the scatterplot. If you place your mouse over one of the green squares at the ends of the line and drag, you can change the slope of the line and move it. You can also use the mouse to move the green dot up and down vertically to change the intercept of the line.

    a. Move the line until you believe your line "best" summarizes the relationship between total mass and skeletal mass for these data. Write down the resulting equation for your line as shown in the applet.

    b. Why do you believe that your line is "best?"

    c. Did all students in your class obtain the same line/equation? How can we decide whether your line provides a better fit to the data than other students' lines? Suggest a criterion for deciding which line "best" summarizes the relationship.

One way to draw the best-fit line is to minimize the vertical distance of the points to the line (these distances are called *residuals*).

5. Would points above the line have a positive or negative residual or is it impossible to tell? What about points below the line?

Key idea
A ***residual*** is the difference between an observed response and the corresponding prediction made by the "best" fitting line (*residual = observed – predicted*). Thus, negative residuals occur when points are below the line and positive residuals occur when points are above the line.

6. Check the **Show Residuals** box to visually represent these residuals for your line on the scatterplot. The applet also reports the sum of the magnitudes of the residuals (**SAE**). SAE stands for "sum of the absolute errors." Errors is another term used for residuals. The acronym indicates we need to take the absolute values of the residuals (or errors) before we add them up.

    Record the SAE value for your line: _____

    What is the best (lowest) SAE in the group? _____

7. It turns out that a more common criterion for determining the "best" line is to instead look at the sum of the *squared* residuals (or errors) (***SSE***). This approach is similar to simply adding up the absolute values of the residuals but is even more strict in not letting individual residuals get too large. Check the **Show Squared Residuals** box to visually represent the squared residual for each observation. Note that we can visually represent the squared residual as the area of a square where each side of the square has length equal to the residual.

a. What is the SSE (sum of squared residuals) for your line?_____

What is the best (lowest) SSE in the group?_____

b. Now continue to adjust your line until you think you have minimized the sum of the squared residuals.

Report your new equation_____

Report your new SSE value_____

What is the best SSE in the class?_____

Key idea
The least squares regression line minimizes the sum of squared residuals.

8. Now check the **Show Regression Line** box to determine and display the equation for the line that actually does minimize (as can be shown using calculus) the sum of the squared residuals.

a. Record the equation of the least squares regression line by indicating the appropriate slope and intercept of the line. Note that we've used variable names in the equation, not generic *x* and *y*. And put a carat ("hat") over the *y* variable name to emphasize that the line gives predicted values of the *y* (response) variable.

$$\widehat{\text{Total Mass}} = \_\_\_\_\_ + \_\_\_\_\_(\text{Skeletal Mass})$$

Notation
The equation of the best fit line is written as $\hat{y} = a + b(x)$, where:
- a is the ***y-intercept***
- b is the ***slope***
- x is a value of the explanatory variable
- $\hat{y}$ is the predicted value for the response variable

b. Did everyone in your class obtain the same equation for the least-squares regression line?

c. Is the slope positive or negative? Explain how the sign of the slope tells you about whether your data display a positive or a negative association.

d. What is the value of the correlation coefficient (Click on the **Correlation coefficient** box to find it.) Explain how the sign of the correlation coefficient tells you about whether your data display a positive or a negative association.

Key idea
For a given data set, the signs (positive or negative) for the correlation coefficient and the slope of the regression line must be the same.

9. Let's investigate what the slope means in the context of total mass and skeletal mass.

    a. Use the least squares regression line to predict the total mass of a bird that has a skeletal mass of 60 grams. (Simply substitute in the value of 60 g for skeletal mass in the equation of the line.)

    b. Use the least squares regression line to predict the total mass of a bird with a skeletal mass of 61 grams.

    c. By how much do your predictions in (a) and (b) differ? Does this number look familiar? Explain.

Key idea
The ***slope*** coefficient of a least squares regression model is interpreted as the predicted change in the response (y) variable for a one-unit change in the explanatory (x) variable.

    d. Interpret the slope in context:

    The slope of the regression line predicting total mass based on skeletal mass is _____, meaning that for every additional _____ gram increase in skeletal mass, the predicted total mass increases by _____ grams.

10. Let's investigate the meaning of the y-intercept in the context of total mass and skeletal mass.

    a. Use the least squares regression line to predict the total mass of a bird that has a skeletal mass of 0 grams.

    b. Your answer to (a) should look familiar. What is this value?

Key idea
The ***y-intercept*** of a regression line is interpreted as the predicted value of the response variable when the explanatory variable has a value of zero (though be wary of extrapolation in interpreting the intercept or other values outside the original data range).

Key idea
Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is known as ***extrapolation*** and can give very misleading predictions.

    c. Explain how your prediction of the total mass of a bird whose skeletal mass is 0 grams is an example of extrapolation.

11. Earlier, we explored the notion of a residual as the vertical distance from a point to a line.

    a. Using the equation for the regression line that you reported in #8, find the residual for the first bird on the list whose skeletal mass is 4.16 grams and total mass is 100.2 grams. (*Hint:* Find this value by taking the actual total mass and subtracting the bird's predicted total mass from the equation.)

b. Is this bird's dot on the scatterplot above or below the line? How does the residual tell you this?

Residuals are helpful for identifying unusual observations. But not all observations of interest have large residuals.

12. Uncheck the **Show Movable Line** box to remove it from the display and check the **Show data options** box and then check the **Move observations** box.

    a. Now click on one bird's point in the scatterplot with skeletal mass near the middle of all skeletal mass values and drag the point up and down (changing the total mass, without changing the skeletal mass too much). You should see that the new regression line is red, and the original regression line will remain in grey. Does the new regression line change much as you change this bird's total mass?

    b. Click on **Revert** to revert the moved point to its original position. Repeat the previous question using the bird with the largest skeletal mass. Drag that observation down vertically so that the total mass is negative. Does this influence/change the regression line more than in part (a) when you chose a skeletal mass near the middle of all skeletal masses? Explain.

Key idea
An observation or set of observations is considered ***influential*** if removing the observation(s) from the data set substantially changes the values of the correlation coefficient and/or the least squares regression equation. Typically, observations that have extreme explanatory variable values (far below or far above $\bar{x}$) are potentially influential. They may not have large residuals, having pulled the line close to them.

Residuals also help us measure how accurate, in general, our predictions are from using the regression line. In particular, we could compare the "prediction errors" from the regression line to the prediction errors if we made no use of the explanatory variable.

13. Press the **Revert** button to reload the original data set and uncheck the **Show data options** box and uncheck the **Show Regression Line** box. Recheck the **Show Movable Line** box to redisplay the blue line. Notice that this line is flat at the mean of the *y* (total mass) values.

    a. Check the **Show Squared Residuals** box (under the **Movable Line** information) to determine the SSE if we were to use the average total mass ($\bar{y}$) as the predicted value for every *x* (skeletal mass). This is a measure of variability in the total mass if we don't take skeletal mass into account. Record this value.

    b. What is the slope of this line?

    c. If the slope of the best fit line is zero, our data shows _____ (positive/negative/no) linear association between the explanatory and response variables.

    d. Now find the SSE value from the regression line by clicking on the **Show Regression Line** box and **Show Squared Residuals** box. How does it compare in size to the SSE from the horizontal line? Why does that make sense?

Coefficient of determination (R^2)

14. A quantity related to the correlation coefficient is called the coefficient of determination, or *R*-squared (R^2).

    a. To measure how much the regression line decreases the "unexplained variability" in the response variable, we can calculate the proportion reduction in SSE with the following formula. Compute the value of R^2.

$$R^2 = \frac{\text{SSE}(\bar{y}) - \text{SSE (regression line)}}{\text{SSE}(\bar{y})}$$

Key idea
The ***coefficient of determination*** (R^2) is the proportion of the total variation in the response variable that is explained by the linear relationship with the explanatory variable.

    b. Complete the following statement: The value of the coefficient of determination (R^2) is

    _____ and this means that _____ % of the

    variation in bird's _____ is explained by the linear relationship with

    their _____.

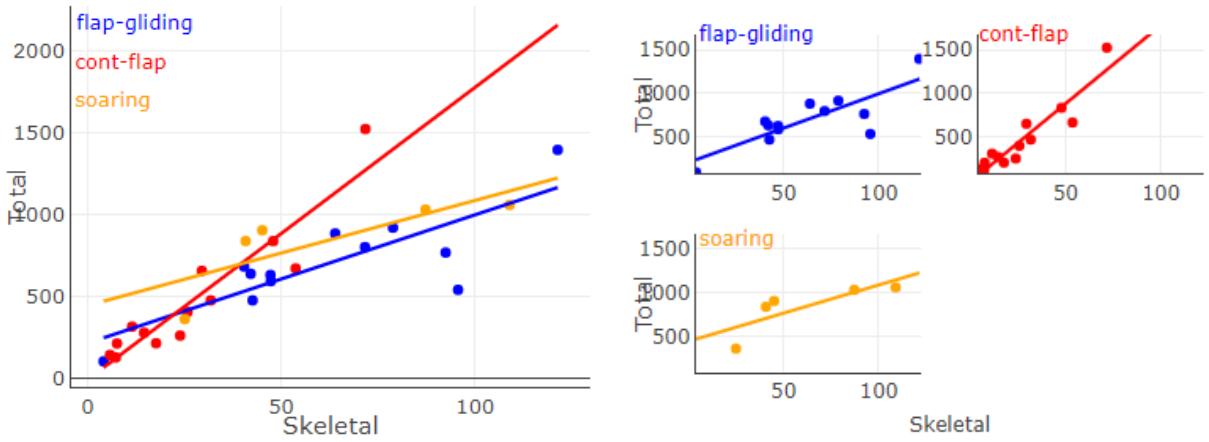
If we think of this horizontal line $\hat{y} = \bar{y}$ as the line we would use if the response variable had no association with the explanatory variable, then we are seeing how much better the actual regression line is than this horizontal line. The best possible line would get rid of all of the residuals and would have an R^2 of 1.

    c. What would a desirable correlation coefficient be? Why? Find the value of the correlation coefficient using the applet and confirm that when you square the correlation coefficient you get the same number as the applet reports for the coefficient of determination (R^2). Unclick **Show Movable Line** and unclick **Show Regression Line**.

Improving explained variation

15. Another variable included in the data set is the type of flight that the bird has (continuous flapping, soaring, flap gliding). Use the **Color by** pull-down menu so the points in the scatterplot are colored by **Flight** type. Do you think the original regression line works just as well for each of these flight types individually or would three different regression lines work better?

The graphs below show the three flight types in different colors and shows the three individual regression lines for each flight type. The graph on the right shows all three flight types together on one graph and the three graphs on the left separate them into individual graphs.



16. Based on the graphs, which flying type do you think has a slope that is most different from the slope of the original regression equation with the combined data?

17. The least squares regression lines for each of the flight types are as follows:

    A. $\widehat{Total\ Mass} = 441.40 + 6.44(Skeletal\ Mass)$
    B. $\widehat{Total\ Mass} = 211.60 + 7.85(Skeletal\ Mass)$
    C. $\widehat{Total\ Mass} = -11.76 + 17.89(Skeletal\ Mass)$

    Use the graph above to match the regression equation (labelled A, B, or C) with the flying type (flap-gliding, continuous-flapping, soaring).

18. Based on the graphs above, which flying type do you think would have the largest R^2 value? Why?

In this case, the R^2 value for the flap-gliding data is 0.653, for the continuous-flapping data it is 0.873, and for the soaring data it is 0.642. So the R^2 value is largest for the continuous-flapping data while the other two are quite similar to the R^2 value for the combined data set.

The three individual regression lines taken together can be considered one model with an R^2 value of 0.813. This means that 81.3% of the variation in total mass can be explained by the linear regression model with skeletal mass and the flight type. This model is an improvement over the one that does not include flight type that had an R^2 of 0.675.

Reference

Martin-Silverstone E, Vincze O, McCann R, Jonsson CHW, Palmer C, Kaiser G, et al. (2015) Exploring the Relationship between Skeletal Mass and Total Body Mass in Birds. PLoS ONE 10(10): e0141794. https://doi.org/10.1371/journal.pone.0141794

Additional reference

For class follow up, use a website for students to guess flapping patterns of birds.
https://virtualexhibits.mos.org/bird-flight-patterns/

Another possible data set to use for this module which has a nice Simpson's paradox in it is the Palmer penguin data. Information for this data set & analyses can be found here:
https://allisonhorst.github.io/palmerpenguins/articles/intro.html