The Stomata Lab Pre-work: Linear Regressions

What the past can tell us about our future – using fossil and modern plants to model atmospheric carbon dioxide.

**Background**

At its simplest, a linear regression attempts to model a relationship between two different variables by fitting a straight line to the data points. That relationship may look something like Graph A or more like Graph B below.

A B

A picture containing line, plot, diagram

Description automatically generated

Mathematically, a linear regression calculates the shortest vertical distance between all graphed data points and a straight line. Typically, statistical programs are used to generate a line that is the closest fit to each of the points.

Biologically, a linear regression helps determines if there is a relationship between two variables, what the strength of that relationship is, and how the relationship may help explain biological phenomena. Specifically, it can inform how much of the differences observed in the dependent variable (*y*) can be explained by the independent variable (*x*).

A straight line can be graphed using the formula:

*y* = *mx* + *b*, where

* ***y***is the dependent variable
* ***x*** is the independent variable
* ***m***is the slope (or steepness) of a line
* ***b*** is the *y*-intercept (where a line crosses the *y*-axis)

However, the equation of a straight line does not indicate how strongly two variables are associated with each other. In the sample linear regressions above, both lines have the same slope, but the data points in the Graph A fit the line closely, while the data points in Graph B are more scattered.

To determine how well data fit to a straight line (or biologically, how strongly the independent variable explains the dependent variable), a value called the **coefficient of determination (*R*2)** is calculated. The formula to calculate *R2* is:

which can also be written as:

(***SS*** is an abbreviation for ***sum of squares***, which is a measure of variance or deviation from the mean or expected value, and the **Σ** symbol represents ***sum***.)

The sum of the vertical distances between observed data points and expected data points used to create a straight line is considered *error*. This error is represented by the dashed lines in the graph below. When calculating *R*2, these errors are summed. The smaller the sum of the error, the better the data points fit to a straight line.

A picture containing line, plot, diagram, parallel

Description automatically generated

*Error* between **observed** data point and **expected** data point

**Expected** data point used to create a straight line

**Observed** data point

To calculate the *R*2 value of data points, the following steps are used:

1. Calculate the **mean** (***ȳ***) of the observed data points.
2. Determine the **total sum of squares** using the following formula:

First, calculate the difference between each observed data point (***yi***) and the mean (***ȳ***). Then, square each difference. Finally, add all the squared differences.

1. Calculate the **expected** data points. These points would fit perfectly to a straight line. The *y*-value of the expected points is represented by ***fi***, and these values are provided in the Practice Example below.
2. Determine the **error of sum squares** using the following formula:

First, calculate the difference between each observed data point (***yi***) and the expected data point (***fi***). Then, square each difference. Finally, add all the squared differences.

1. Using the **total sum of squares** and the **error of sum squares** ,  calculate the **coefficient of determination (*R*2)** using the following formula:
2. Once *R2* has been calculated, both mathematical and biological interpretations can be made. *R*2 values range from 0 to 1.

Mathematically, the closer *R*2 is to 1, the closer the observed data points fit to a straight line. Therefore, when *R*2 is close to 1, there is minimal error between the observed data points and the expected data points represented by a line.

Biologically, the closer *R*2 is to 1, the more confidently the dependent variable (*y*) can be predicted by the independent variable (*x*). Different fields of study (e.g., social sciences, biology, physics, etc.) have different minimum values for *R*2 necessary to suggest that a relationship is “good” or “strong.”

**Practice Example**

Water conservation is essential to all terrestrial organisms. Animals and plants that live in drier environments have adaptations to conserve water, and these adaptations are less likely to be found in animals with more consistent access to fresh water. Some of these adaptations may be behavioral, but many can also be physiological. For example, in mammals, kidneys are responsible for removing certain wastes from the blood and for reabsorbing water when the animal is dehydrated.

This is why diuretics such as caffeine increase urination. They block the re-absorption of water in the kidney, meaning more water is lost through urine.

To better understand the importance of kidneys to water conservation, a relationship could be established between the **relative size of kidneys** in closely related species and the **different amounts of rainfall** received by the environments in which these species live. A study by Davis et al. (2006) explores this relationship. Some of these data are included below, and the graph axes values have been removed for simplicity.

A picture containing line, plot

Description automatically generated

Sample data points for relative rainfall (Column A) and relative kidney size (Column B) are provided in Table 1 on the next page. Use the steps provided (1 – 6) to complete Table 1 and calculate the **coefficient of determination (*R*2)** of the sample data.

In this example, relative rainfall of the environment is the independent variable (*x*), and relative kidney size is the dependent variable (*y*). Mathematically, *R*2 describes how much error there is between the observed and expected kidney size. Biologically, describes how much of the differences in kidney size between species can be explained by the amount of rainfall in their environments.

Table 1. Calculation of *R2* using sample data points for relative rainfall and relative kidney size.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A | B | C | D | E | F | G | H |
| **Rainfall**  **(mm)** | **Observed Kidney Size**  **(cm3)** | **Mean Observed Kidney Size**  **(cm3)** | **Difference between Observed and Mean**  **(cm3)** | **Difference between Observed and Mean Squared** | **Expected Kidney Size**  **(cm3)** | **Difference between Observed and Expected**  **(cm3)** | **Difference between Observed and Expected Squared** |
|  |  |  |  |  |  |  |  |
| 100 | 56 |  |  |  | 50 |  |  |
| 150 | 37 |  |  | 40 |  |  |
| 200 | 29 |  |  | 30 |  |  |
| 250 | 18 |  |  | 20 |  |  |
| 300 | 12 |  |  | 10 |  |  |
|  |  |  | |  |  | |  |

1. Using the Observed Kidney Sizes in Column B, calculate the **mean** Observed Kidney Size, and write this value in Column C.
2. Determine **total sum of squares** using the following formula:
3. First, calculate the difference between each Observed Kidney Size (Column B) and the mean Observed Kidney Size (Column C). Write these differences in Column D.
4. Then, square each value in Column D, and write these values in Column E.
5. Finally, add all values in Column E to calculate *SStotal*. Write this sum where indicated at the bottom of Column E.
6. The expected Kidney Sizes have already been calculated and are included in Column F. These data points would fit perfectly to a straight line.
7. Determine the **error of sum squares** using the following formula:
8. First, calculate the difference between each Observed Kidney Size (Column B) and the Expected Kidney Size (Column F). Write these differences in Column G.
9. Then, square each value in Column G, and write these values in Column H.
10. Finally, add all values in Column H to calculate *SSerror*. Write this sum where indicated at the bottom of Column H.
11. Calculate the **coefficient of determination (*R*2)** using the following formula and by transferring the values calculated for *SStotal* and *SSerror* from Table 1 to the boxes below.

+

1. Consider how the value of *R*2 would be interpreted mathematically and biologically.

**Notes:**

* The **coefficient of determination (*R*2)** can also be calculated as:

This equation and the one used in the pre-work would compute the same value for *R*2.

* The Pearson correlation coefficient (*r*) is calculated by taking the square root of *R*2.
* Diaz, G. B., Ojeda, R. A., & Rezende, E. L. (2006). Renal morphology, phylogenetic history and adaptation of South American hystricognath rodents. *Functional Ecology, 20*, 5-13. <https://doi.org/10.1111/j.1365-2435.2006.01144.x>

**Conclusion Questions**

1. Based on the *R2* value calculated in the Practice Example, how well do the observed data points fit a straight line? Justify your answer.
2. Mathematically, explain what *R*2 means as it relates to the Practice Example.
3. Biologically, explain what *R*2 means as it relates to the Practice Example.