# Bird Banding: MAPS Data Analysis

**PART 1 Visualizing Data with Scatter Plots**

**Goals:**

- Examine the relationship between survival and reproductive rates in birds from the three main ecoregions of Minnesota.
- Create a scatterplot to visualize the relationship between these two quantitative variables
- Qualitatively describe the characteristics of the scatterplot describing the direction, form and strength.

Part 2 **Numerical summaries and Correlation**

**Goals**

- Calculate a correlation coefficient
- Recognize that a correlation coefficient of 0 means that there is no linear association between the two variables and that a correlation coefficient of –1 or 1 means that the scatterplot is exactly a straight line.
- Create a null and alternate hypothesis for explaining the relationship seen between reproduction and survival
- Test your hypothesis and estimate a p-value using simulations

**PART 3 Least Squares Regression**

**Goals**

- Summarize the scatterplot by finding the best-fit (least squares regression) line.
- Understand what residuals are and their role in finding a best-fit line
- Interpret 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.
- Hypothesize why the observed values do not match the predicted values in many cases
- Find and interpret the coefficient of determination ($R^2$)
- Estimate a p-value using simulations

**Part 4 – Using R**

**Goals**

- Create a scatterplot in R and adjust the X and Y axes,
- Calculate the correlation coefficient in R

- Conduct a Pearson Correlation Test to calculate a p-value for the correlation coefficient in R
- Create a linear model in R
- Plot a regression line in R
- Find the $R^2$ and p-values for the linear model in R

**Background:**

For a population to be sustained, a pair of sexually reproducing organisms needs to, on average, produce two or more offspring that survive to breed themselves during their life. There are two basic reproductive strategies that species can use to achieve this. On one end of the spectrum, some species maximize the number of young produced but provide little care and resources to any individual. This is referred to as R or "reproductive" selection. Other species will produce few offspring but invest more in caring for each. This is known as K or "care" selection.

In this lab, we will use data collected from bird species that breed in Minnesota to visualize, describe, and quantify the relationship between reproduction and survival in birds and see if the patterns we see fit what we might expect given the reproductive strategies described above.

**Study System and Data:**
Studies of breeding birds are a good way to understand the dynamics between rates of survival and reproduction. The Monitoring Avian Productivity and Survival (MAPS; https://www.birdpop.org/pages/maps.php) program uses bird banding techniques to collect bird population and demographic data, including information on the number of young relative to the number of adults in a population and annual probability of survival. Collaborators from 1200 bird banding stations across North America have contributed data from 2.5 million birds. With the information collected, it is possible to determine the relationship between survival and reproduction rates in individual species in different regions of North America.
You will have hopefully been able to join us for one of our bird banding demonstrations in the natural lands to learn how birds are captured and measured; and how age, sex and reproductive status are determined. However, that won't be required for this exercise.

The MAPS data that I am providing for this assignment has been accessed from a much larger dataset that can be found here:

https://tinyurl.com/Bio150Maps

I've pared down the data we're working with today for this exercise, but a wide variety of data is available (feel free to explore what is available!) For our purposes, we're considering two variables from the MAPS results. Annual adult survival probability and the reproductive index, which is a relative measure of young produced per adult. We will be using these data from birds sampled from the three major ecoregions of Minnesota: prairie pothole, boreal forest, and hardwood forest.

**Part 1: Visualizing Data with Scatterplots:**
We will begin by exploring these data using the **Regression** applet, which can be found here.

**You can download the specific data you will be using here.**

Think about questions that are posed about each step and write your answers down for those labeled with **'Question'**

**Scatterplots:**
To use the data:
1) Open the spreadsheet you just downloaded
2) Highlight all rows and columns
3) Copy the data
4) Open the **Regression** applet and click "Clear" under the data that currently exists there.
5) Click inside the empty box and paste the data into the applet. Click "Use Data".
You should now see a sample size of "$n = 46$" below your dataset. The variable *survival* is the probability of an adult individual to survive into the next breeding season, and the variable *reproduction* is the reproductive index - a relative measure of young produced per adult each breeding season. We will consider the reproduction index to be the explanatory variable and the response to be survival probability. You can reverse those roles by toggling the **Explanatory, Response** button or using the pull-down menus next to **Response Variable** and **Explanatory Variable.**

**Key Idea**

A ***scatterplot*** is a graph showing a dot for each observational unit, where the location of the dot indicates the values of the observational unit for both the explanatory and response variables. Typically, the explanatory variable is placed on the $x$-axis and the response variable is placed on the $y$-axis.

What is an explanatory variable? What is a response variable?

**Question 1-1:** Why are we considering reproduction to be the explanatory variable in this system? (Hint: Think about the lifecycle of salmon).

What if we were considering the lifetime reproduction and survival in these populations instead of year-to-year reproduction and survival? Would that change how we think about the explanatory and response variable?

**Qualitative Summaries**
When describing a scatterplot, we look for three aspects of association: direction, form, and strength. The ***direction*** of association between two quantitative variables is either positive or

negative, depending on whether the response variable tends to increase (positive association) or tends to decrease (negative association) as the explanatory variable increases.

Is the observed association between reproduction and survival (the **direction**) positive or negative?

**Question 1-2:** Does this match your expectations of how reproduction and survival would be related? Why/why not?

The ***form*** of association between two quantitative variables is described by indicating whether a straight line would do a reasonable job summarizing the overall pattern in the data or some other pattern such as a curve would be better. It is important to note that, especially when the sample size is small, you don't want to let one or two points on the scatterplot change your interpretation of whether or not the form of association is linear. In general, assume that the form is linear unless there is compelling (strong) evidence in the scatterplot that the form is not linear. Some examples of non-linear relationships that might be observed include quadratic or logarithmic relationships.

Does the association between reproduction and survival in these species appear to be linear or is there strong evidence (over many observational units) suggesting the relationship is nonlinear? i.e. what is the *form?*

In describing the ***strength*** of association revealed in a scatterplot, we see how closely the points follow the form: that is, how closely do the points follow a straight line or curve. If all of the points fall pretty close to a straight line or curve, we say the association is strong. Weak associations will show little pattern in the scatterplot, and moderate associations will be somewhere in the middle.

In your opinion, would you say that the association between reproduction and survival appears to be strong, moderate, or weak? i.e. what is the *strength*?

**Question 1-3**: How does this match expectations you'd have regarding the relationship between reproduction and survival?

Part 2 **Numerical summaries and Correlation**

Describing the direction, form, and strength of association based on a scatterplot, along with investigating unusual observations, is an important first step in summarizing the relationship between two quantitative variables. We can also use a statistic to summarize the association. One of the statistics most commonly used for this purpose is the correlation coefficient, which measures the strength and direction of the *linear* association.

The ***correlation coefficient***, often denoted by the symbol ***r***, is a single number that takes a value between –1 and 1, inclusive. Negative values of *r* indicate a negative linear association, whereas positive values of *r* indicate a positive linear association. The stronger the linear association between the two variables, the closer the value of the correlation coefficient will be to either –1 or 1, whereas weaker linear associations will have correlation coefficient values closer to 0. Moderate linear associations will typically have correlation coefficients in the range of 0.30 to 0.70 or –0.30 to –0.70.

> Before you look at the applet ask: will the value of the correlation coefficient for the reproduction survival data be negative or positive? Why?
>
> Without using the applet, estimate the correlation coefficient between reproduction and survival based on the scatterplot.

Check the **Correlation coefficient** box in the applet to reveal the actual value of the correlation coefficient.

> **Question 2-1:** What is the calculated value of the correlation coefficient?

It is important to consider any ***unusual observations*** that do not follow the overall pattern. The correlation coefficient is sensitive to unusual observations. Removing an influential observation can substantially change the value of the correlation coefficient. Click on one of the more unusual observations and press the delete button and see how it changes in the correlation coefficient. Try this with a few points and see how the correlation coefficient changes.

Check the box for **Show data options** then check the box for **Move observations**. Now put the cursor on the point farthest to the right in the graph and slowly slide it upwards.

> What happens to the value of the correlation coefficient as you slide the point up? Can you slide it up far enough to changes the direction of the correlation?

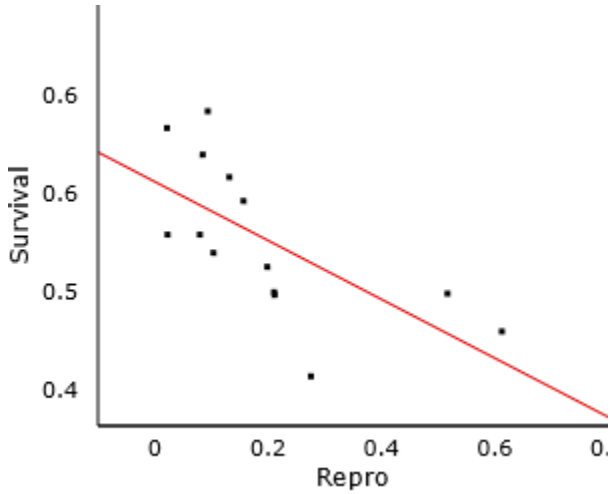
Would you consider this new moved point an influential observation?

**Question 2-2:** How should highly influential observations be handled?

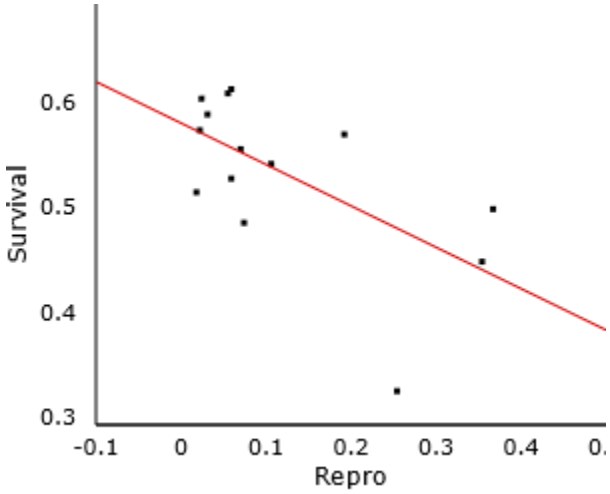
Press the **Revert** button to revert the data to its original form. You will notice that there is a third variable of region included in the dataset. Under the **Change variable selections**, in the pull-down menu for **Color by**, select region. This will include a third variable (this time categorical) into the graph by making each observation from each ecosystem a different color.

The three graphs below show separate scatterplots for the three ecoregions to make it easier to directly compare associations.

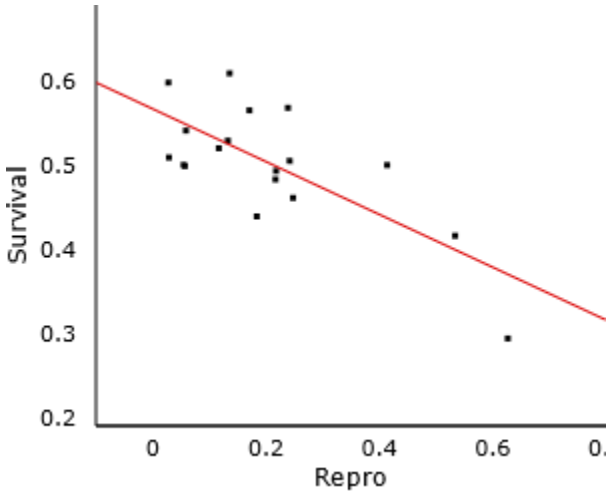
**Boreal**



**Prairie**




**Hardwood**




How does the relationship between reproduction and survival compare between birds in the three habitat types?

**Question 2-3:** Does this match what you'd expect based on what you know about reproductive strategies? Why or why not?

Now we want to test whether the correlation observed between reproduction and survival represents a meaningful, non-random relationship. This is known as hypothesis testing – the 'null hypothesis' in any system is that there is no underlying difference or relationship in the data being examined and an alternative hypothesis is that there is some difference or relationship that did not occur at random. To conduct these hypothesis tests, we will start by using simulations in the Regression applet to visualize how likely a correlation of similar strength to the one we observed might occur at random.

.

We will use the Regression applet to generate a large number of simulated results with the same sample size as our data that assumes no underlying association (any *y*-value in the dataset can be paired with any *x*-value), calculating the correlation coefficient for each one, and seeing how often we obtain a correlation coefficient as or more extreme (as far from zero) as what we have observed in our data.

**Question 2-4**: When considering this relationship, what is our null hypothesis? What is our alternative hypothesis?

Continuing with the **Regression** applet, check the **Show Shuffle Options** box and select **Correlation** from the pull-down menu by **Choose statistic**. Then press **Shuffle Y-values** to simulate one result assuming there is no association between the two variables. Shuffle Y-values several more times and look at what the scatter plots for the randomized data look like and look at the histogram on the right to see what correlation coefficient values are being produced. Now increase the number of shuffles to 1000 and look at the frequency of correlation values.

Next to the **Count Shuffles** box choose **Beyond** from the pull-down menu. Specify the observed correlation coefficient for our data and press **Count**. This represents an approximate **p-value** (see Key Idea box).

**Key Idea**

A ***p-value*** a statistical calculation that assesses whether the observations you have made are likely to have occurred at random. A low ***p-value*** (in biology, we often use a threshold $< 0.05$) means that there is a low likelihood that your observations were due to chance and a high likelihood that there is a real trend. A high ***p-value*** means the opposite – that it is unlikely that a real trend exists.

What proportion, of the ~1,000 simulated random results produced a correlation coefficient as extreme (as far from zero in either direction) as our observed value?

**Question 2-5:** How does this value relate to a p-value?

Now try the same simulation using just data from the boreal region. To use the boreal-only data, you will need to copy just the boreal data into the applet from this spreadsheet. What difference do you notice about the estimated p-value and correlation coefficient? What is causing this difference?

What conclusion would you draw from this approximate p-value?

Is it possible to have a very high correlation coefficient accompanied by a high p-value or a very low p-value with a low correlation?

**PART 3 Linear Regression**

Now that we've described the correlation between survival and reproduction, we want to more closely quantify the nature of that relationship and use linear regression to predict survival based on reproduction.

Go back to the **Regression** applet and check the **Show Movable Line** box to add a blue line to the scatterplot. If you click and drag the green boxes on the line, you can change the slope of the line or the intercept.

Move the line up and down and change it's slope until it looks like it follows the overall trend of the points on the scatterplot. Move the line until you find what seems like it "best" describes the relationship between survival and reproduction. (This will just be an educated guess on your part.)

**Question 3-1**: What is the equation of your "best" line? Why do you believe your line to be the best?

One way to objectively calculate a best fit line is to minimize the total vertical distance of the points to the line (these distances are called *residuals*). That is to say, on average, all the points are as close to the line as possible.

Check the **Show Residuals** box to visually represent these residuals (also referred to as error) for your line on the scatterplot. The applet reports these as the sum of the magnitudes of the residuals (also called **SAE** - sum of the absolute errors).

Next check the **Show Squared Residuals** or the *sum of the squared residual* (aka **SSE** sum of squared error). This visually represents that the squared residual is the area of a square where each side has a length equal to the residual. SSE is most commonly reported, because this approach is stricter in not letting individual residuals get too large.

> **Question 3-2:** What is the SSE value for your original line? What is the best (lowest) SSE in the class?

Looking at the SAE and SSE, continue to adjust your line until you think you have minimized the sum of the squared residuals. Now check the **Show Regression Line** box to show the equation for the exact line that minimizes the sum of the squared residuals. This is known as the least squares regression line or the best fit line.

Record the equation of the least squares regression line below using the 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.

$$\hat{Survival} = \_\_\_\_\_ + \_\_\_\_\_(Reproductive\ Index)$$

In theory, you can use this equation to predict the value of one variable based on the other.

**Notation**

**The equation of the best fit line (aka least squares regression 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**

**Coefficient of determination ($R^2$)**

So, we now have a line that best fits our data. Now the question is: how good is the best we can do? How well does this equation actually fit our observations? The most commonly used measure for this is called the coefficient of determination, or $R$-squared ($R^2$). This number is closely related to the correlation coefficient and in a system with only two variables (one explanatory and one response), it is the square of the correlation coefficient. The $R^2$ is a preferable metric, because it can also be applied to a system that includes more than one explanatory variable.

**Key idea**
The ***coefficient of determination*** ($R^2$) is the proportion of the total variation in the response variable that is accounted for by changes in the explanatory variable(s).

A horizontal line and random distribution of points would have an $R^2 = 0$. The best possible line would get rid of all of the residuals and would have an $R^2$ of 1. There is no hard and fast rule in science of what a "good" $R^2$ value is. In some systems, being able to describe 20% of the variation in a population is extremely useful.

**Draw inferences – Estimating a P-value.**
You will have found an association between survival and reproduction. The question, however, is how probable is it that an equally strong association will occur by chance?

We will use simulations to model the null hypothesis, that there is no relationship between reproductive index and survival. That is, any response value is just as likely to be paired with any explanatory variable value. Check the **Show Shuffle Options** box, select the **Slope** from the **Choose statistic** pull-down menu and press **Shuffle Y-values** to shuffle the response variable values (survival), reassigning them at random to the explanatory variable values (reproductive index).

How does the regression line of the randomized data (blue) relate to the regression line of the observed data (red)? What is the value of the slope for the shuffled data?

Now do at least 1,000 shuffles and describe the pattern of the 1,000 regression lines across the different shuffles.

Use the applet to count how many of the shuffled slopes are at least as extreme as the observed slope. What is your estimated p-value?

**Part 4 – Correlations and Linear Regression Model in R – See R Notebook**

r.stolaf.edu

**Question 4-1:** How does calculating correlation coefficients and regression equations differ between R and the Applet? When would you use each analytical method?

**Overall Questions:**

1) If a population needs to balance reproduction with survival in order to be sustained, why aren't these indicators perfectly correlated across all species across all regions? Think of some meaningful sources of variation in addition to possible errors made during data collection.

2) Why don't/can't species maximize both reproduction and survival?

3) If a species is an outlier above or below the expected ratio of survival to reproduction, how might we expect its population to be changing? (Assuming we confirm that our estimations of reproduction and survival are largely accurate.)

4) Look at some more of the data available on the MAPS website (https://tinyurl.com/Bio150Maps). What further questions could you ask about evolution from the information available?

5) What are some advantages to using R vs using excel or online-based statistical programs?