MODELING CHANGES IN PROBABILITY OF PRESENCE OF MYOTIS LUCIFUGUS DUE TO WNS: AN INTRODUCTION TO LOGISTIC REGRESSION IN R

Mary Roseblock | Virginia Commonwealth University

Learning Objectives

Environmental: Students will learn more about the wildlife disease known as white nose syndrome (WNS) and why it causes high mortality in little brown bats (*Myotis lucifugus*). Students will be able to list the important ecosystem services bats provide, and they will be able to discuss the importance of conserving this species.

Conservation: Little brown bats are currently under review for listing under the ESA (Endangered Species Act). Students will learn about the broader protections provided under the ESA. Students can decide if they support listing this species as federally endangered.

Data Analysis: Using data from a previous study, a logistic regression model will be fitted to evaluate the relationship between WNS and probability of presence of little brown bats. Students will learn about logistic regression models, evaluate the fit of the model, interpret the coefficients, and extract the probability from the logistic regression equation.

Prerequisites

This lesson requires basic knowledge of the R programming language and familiarity with linear regression.

ENVIRONMENTAL- OBJECTIVE ONE

WNS is caused by the fungus *Pseudogymnoascus destructans* (PD). The first documented case of WNS in the United States occurred in a popular tourist cave in New York state. It quickly spread throughout the Appalachian Mountain range aided by both bat migration and cave tourism. The fungus likely arrived in North America from Europe or Asia; however, the disease does not cause mass mortality in European or Asian bats (Hayes, 2012).

PD, unlike many other fungi, is cold-loving and thrives at temperatures ranging from 4 to 20 degrees Celsius. PD's ability to thrive at low temperatures has made it an incredibly destructive pathogen targeting bats that hibernate communally in damp, cold caves during the winter. Not every species is affected equally - size, metabolism, and even behavior all play a role in survival outcomes.

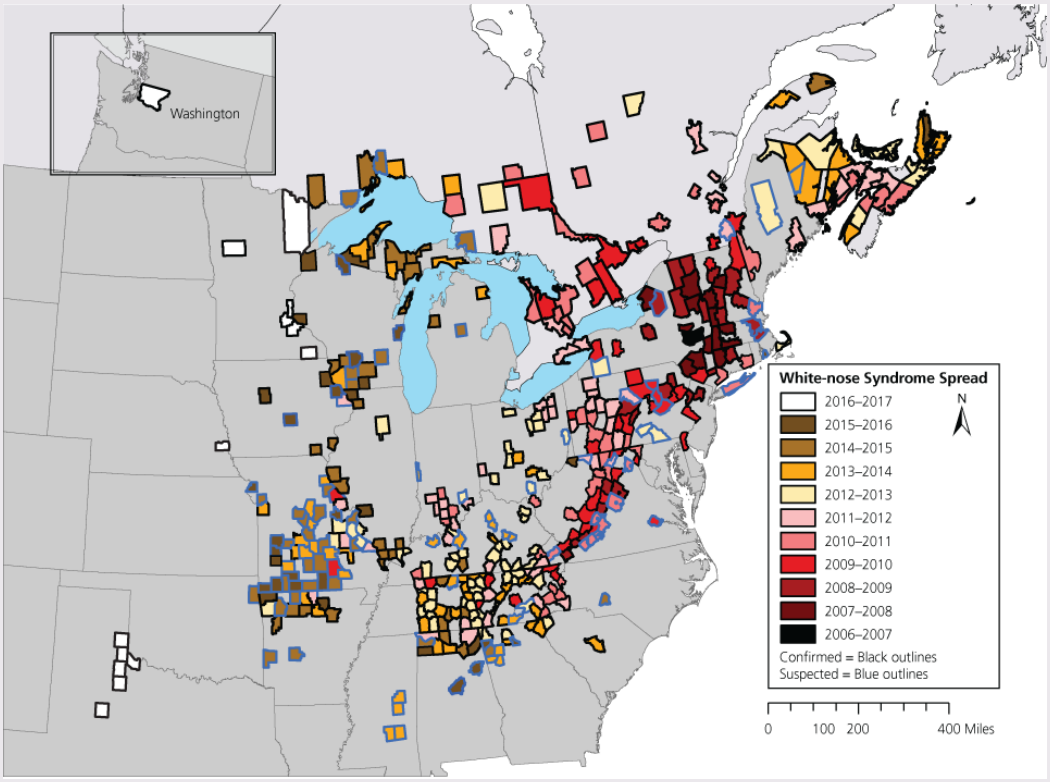
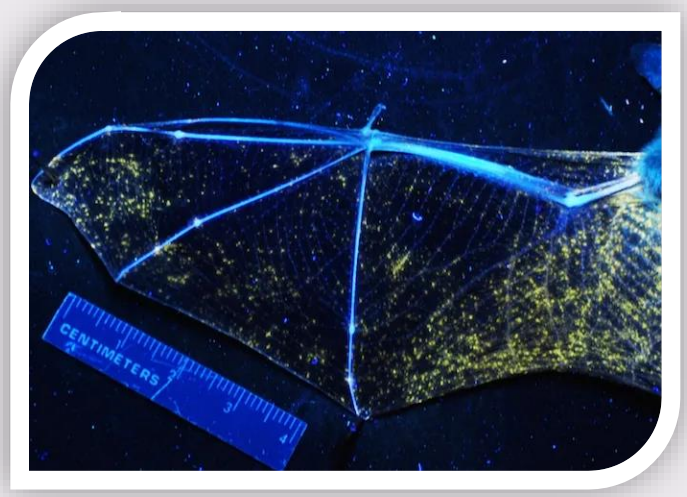


Figure 1: Spread of WNS from 2006-2017 (Wynne, 2017).

Three species comprise the bulk of the overall loss of individuals: the northern long-eared, little brown, and tri-colored bats. These species declined by 90% from 2003 to 2013.



The fungus invades the skin tissue of hibernating bats and typically colonizes and extends its fungal hyphae throughout the wing; the colonization of the wing membrane and subsequent damage disrupts the water balance of hibernating bats. Bats suffering from PD have a characteristic white muzzle (Beck et al., 2017). Although the exact mechanism that causes mortality from WNS is not fully understood, it is hypothesized that the disease causes bats to exit hibernation early which leads to a cascade of physiological effects. Most notable is the rapid depletion in energy stores. Bats that exit hibernation early because of WNS are faced with decreased resource availability and reduced flight capacity as a result of wing damage (Verant et al., 2014). If you would like to learn more about WNS, watch this video: https://www.youtube.com/watch?v=DqvvnEObh4A

Disease progression varies by species, with some species inherently better-suited to recover from the disease. For this lesson, we will be focusing on *Myotis lucifugus* (little brown bats). This once prolific species is particularly vulnerable to WNS for the following reasons:

- Their metabolisms do not allow them to readily compensate for evaporative water-loss (Beck et al., 2017).
- They are small bodied and have low body fat.
- They prefer to hibernate in very humid caves. Species with a preference for drier caves appear to be less impacted (Manness, 2016).

Why should we care about little brown bats? Little brown bats provide important ecosystem services. They eat lots of bugs, especially nuisance insects like mosquitoes! Little brown bats can eat 70% of their bodyweight in insects. This controls the population of many insects and potentially slows the spread of vector

borne diseases. Additionally, they protect crops from pests and reduce the overall need for pesticides (USFWS, 2023). It's estimated that bats save farmers $3 billion per year in the United States (USGS, 2011).

They are food for other species, and their decline has cascading effects on trophic levels. Raptors and raccoons depend on bats for calories, and a decline in little brown bats means there will be less available food for the next trophic level. Little brown bats are potential bioindicators and are sensitive to climate change and disease (Shively, 2017). Conserving this species would ensure it continues to provide these important services.

Stop and Read Section 1 of the attached Johnson et al. (2021) paper.

1. What year was WNS first discovered in West Virginia? 2009
2. What did the authors expect to see in the count trends for *M. lucifugus* and *P. subflavus*? They expected these two species to decline steeply.

CONSERVATION-OBJECTIVE 2

Little brown bats are currently under review by the USFWS to determine if they are eligible for listing under the Endangered Species Act (ESA). A handful of states, including Virginia, consider little brown bats to be state endangered; however, listing under the ESA will provide broader federal protection for the species. The ESA protects species by restricting development or clearing of final critical habitat, prohibiting federal action that would jeopardize the species' existence, and making it unlawful to take a species without a permit. The ESA also requires dedicated recovery plans to assist population recovery (Taylor, Suckling, & Rachlinski, 2005).

Stop and Read Section 4 (Management Implications) of the Johnson et al. (2021) paper and answer the following questions.

3. What do the authors conclude is critical to recovery? Habitat management focused on helping WNS-affected bats achieve energy balance upon emergence from hibernacula could be critical.
4. Endangered Species Act help with recovery? This question is intended to be open ended, but students should use information from the ESA section describing habitat protection. A primary goal of the ESA is

protecting final critical habitat. Protecting critical habitats and their resources could help bats recovering from Pd.

DATA ANALYSIS-OBJECTIVE 3

The focal paper makes use of existing summer surveys conducted by the Monongahela Forest service from 2003 to 2019. The data also includes information on body condition and reproductive state. For this lesson, we will be focusing on the count data only. The authors predicted that summer count trends would mimic the same steep declines as hibernacula counts for WNS affected species. The authors used a zero-inflated Poisson generalized additive model for their analysis. We will recreate the presence trend from Figure 3(a) by simplifying and using a type of GLM called a logistic regression. Since the analysis is simplified, and we are only focusing on probability of presence, only sections 1, 2.1, 2.2, 3.1, and 4 of the focal paper need to be read. It's important to read the relevant sections of the focal paper before continuing to the data analysis section.

Stop and Read Section 2 (Study Area and Data Collection) and Section 3 (Presence and Relative Abundance) of Johnson et al. (2021) paper and answer the following questions.

5. How many sites were surveyed? 209
6. Were the sites selected randomly? Were the counts independent? The sites were not randomly chosen. This is likely due to the fact that it was difficult to find bats, and the surveyors went to locations where bats were known to frequent. The counts of the bats are independent.
7. How many species were heavily impacted by WNS? 3 of the 8 species surveyed were heavily impacted.
8. What was the approximate probability of presence in 2019 for *Myotis lucifugus*? Based on figure 3(a), it's around 0.05.

DATA ANALYSIS-OBJECTIVE THREE

WHAT IS A GLM?

GLMs allow us to model data that are not normally distributed (Hector, 2015). Non-normal distribution is common among natural science data since they generally consist of counts. GLMs have three components: (1) the random component, (2) linear predictor, and (3) the link function.

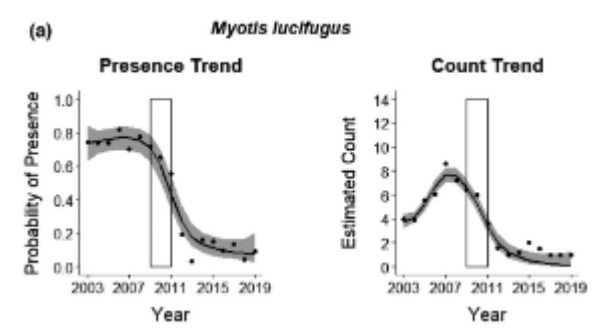
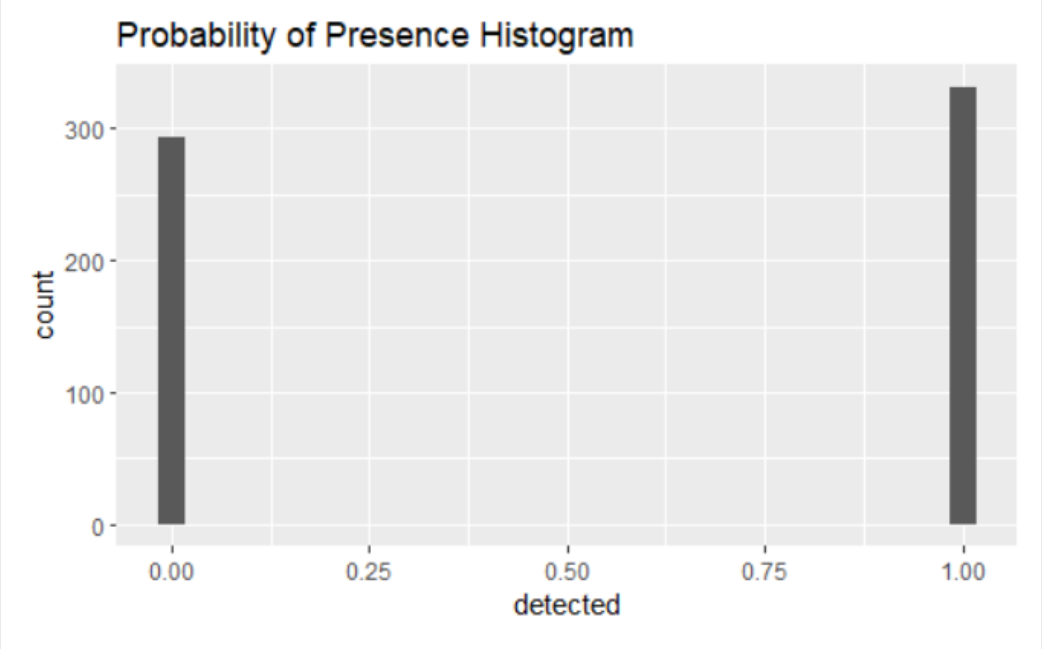


Figure 2: Figure (3a) from the focal paper that we will be recreating in-part (Johnson, 2021). It shows the probability of presence for little brown bats and the count trend.

1. The random component or "family" specifies the distribution of the response variable. We can determine family by evaluating a histogram of our data or by considering the nature of our data. For example, Poisson and negative binomial are common distributions for count data.
2. The linear predictor is the set of independent variables that will be used to predict the dependent variable. 3
3. The link function maps a nonlinear relationship to a linear one so that a linear model can be fit. It does so through a mathematical transformation, and the transformation used depends on which link function is chosen.

ABOUT LOGISTIC REGRESSION

Logistic regression predicts whether something is true/false or yes/no. It is a type of generalized linear model that can be used with binary data. Since the data are binary, the binomial family of distribution is used. The response variable will be bounded by 0-1. Let's look at a histogram of our little brown bat data to confirm this.



Logistic regression uses the logit link function, which means the response variable Y is transformed to log odds (PSU, 2023). This transformation allows us to interpret the non-linear relationship between our response variable and our predictor variable in a linear way. The logistic regression equation is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + B_1(X)$$

You can see the equation is pretty similar to the equation for linear regression! B_0 is our Y intercept coefficient, B_1 is our slope coefficient, and X is our predictor variable. Since our Y values are "logits", interpreting our coefficients won't be as intuitive as with linear regression, and we will be interpreting the change in the odds of Y for every 1 unit increase of X. You will also notice in the equation that the odds portion is written in terms of probability. Let's first discuss what odds are and how they are related to probability.

Odds are the number of times an event occurred divided by the number of times the event did not occur. Probabilities measure the number of times an event occurred divided by all possible outcomes. For example, the probability of getting tails when you flip a coin is 1/2 or 50% because there are two possible outcomes. The odds of getting tails when you flip a coin is 1/1 because there is one chance tails will occur and

one chance that it will not occur. Alternatively, you can find the odds given the probability by using a simple formula: probability/(1-probability). You will notice from the logistic regression equation that the log odds are written in terms of p or probability. We can find the probability of presence for a given year by re-arranging the equation to solve for probability.

FITTING THE MODEL, RUNNING DIAGNOSTICS, AND INTERPRETING THE COEFFICIENTS

This portion of the lesson is intended to accompany the R markdown file, and it contains additional information describing what is being done in R. You may follow along if you wish.

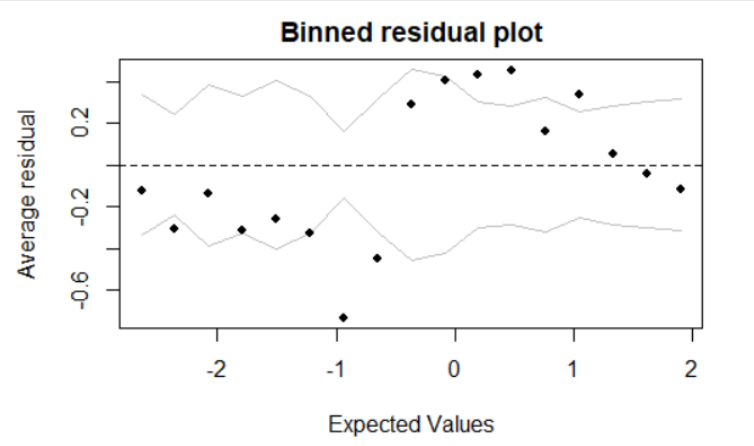
To fit a logistic regression model, we can still use the GLM function in base R. Setting the family to binomial and link function to logit tells R to fit a logistic regression.

```
glm(formula = detected ~ Year, family = binomial(link = "logit"),
    data = lbb)
```

Before we interpret the coefficients of the model, we should run some diagnostics to make sure it's a good fit. The first thing we can check are the deviance residuals from the model summary. Deviance residuals measure how much the probabilities estimated from our model differ from the observed proportions of successes. Min and max should have roughly the same absolute value. 1Q and 3Q should also have similar absolute values (Ford, 2022). The deviance residuals indicate the model is a pretty good fit.

```
Call:
glm(formula = detected ~ Year, family = binomial(link = "logit"),
    data = lbb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0189  -0.8131   0.5284   0.8745   2.3270
```

Next, we can look at a binned residual plot that looks at average residuals as a function of expected values. With most generalized linear models, the standard diagnostic plots created by the plot function can be used. However, the constrained nature of binary data makes these plots unhelpful. Instead, we can look at binned residual plots to visually assess the overall fit of our model (Hector, 2021). The grey lines represent ±2 standard errors. Approximately 95% of the points should fall within the grey lines. We can see the majority of our points fall within the grey lines, but it's not 95%. This could be due to the fact that binning creates 17 data points (binned by year), and the plot doesn't work as well when there aren't a lot of bins (Hector, 2021). We will move forward with interpreting the coefficients from summary output with this in mind.



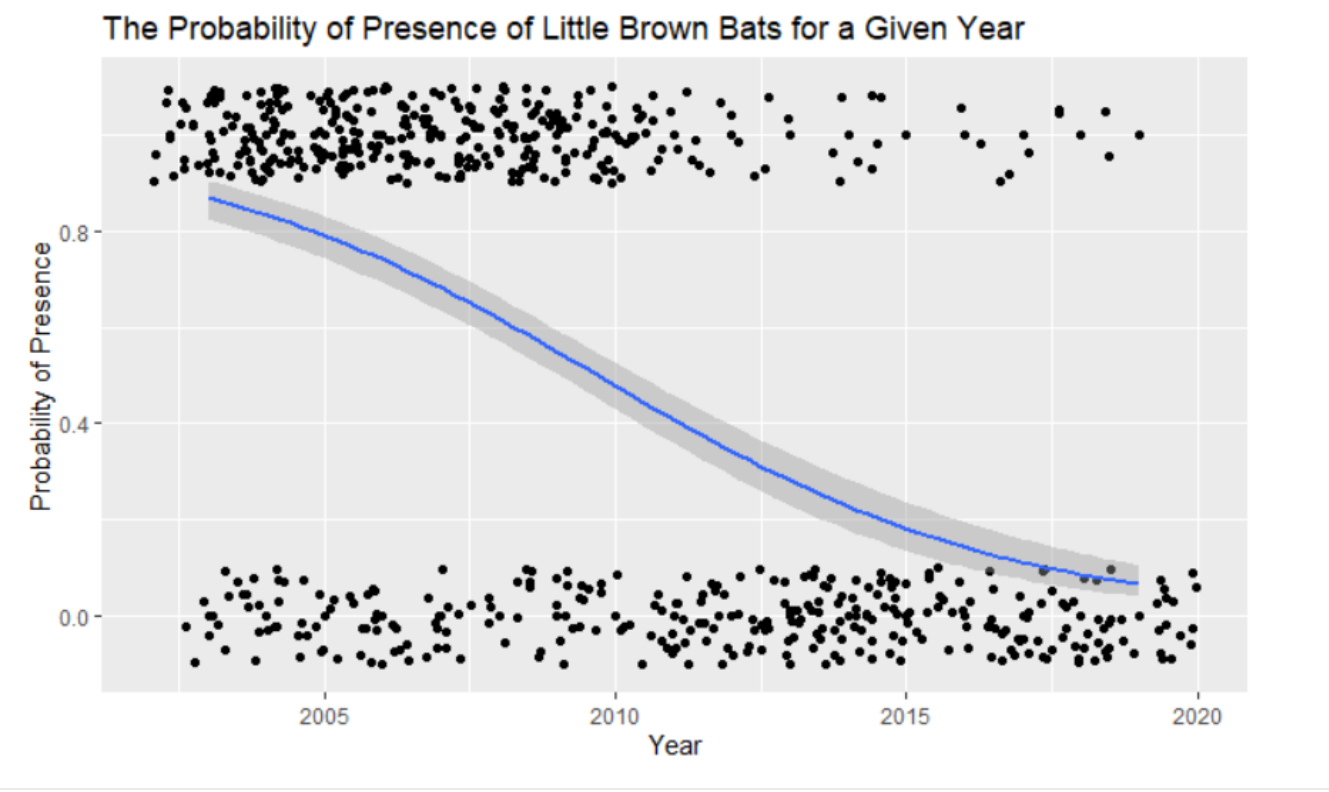

```
Call:
glm(formula = detected ~ Year, family = binomial(link = "logit"),
    data = lbb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0189  -0.8131   0.5284   0.8745   2.3270

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 569.82594   48.91799   11.65   <2e-16
Year         -0.28354    0.02435  -11.64   <2e-16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 862.73  on 623  degrees of freedom
Residual deviance: 670.50  on 622  degrees of freedom
AIC: 674.5

Number of Fisher Scoring iterations: 4
```

The y-intercept tells us what the value of y should be when x is equal to zero. Because the year 0 is outside of our data set, this value is not useful on its own. However, it is fundamental to our logistic equation. We will need it to solve for probability of presence.

Our p-value is less than 0.05, so the model suggests that there is a statistically significant association between the predictor variable and response variable.

This is our slope coefficient in log odds. This shows that for every 1 unit increase in x, there is 0.2978 decrease in the log odds of y.

Here's what the plot of our model looks like:



Since log odds are not very intuitive, we will find the odds first and use these values to calculate probability of presence for a given year. The odds ratio can be found by exponentiating the log odds and subtracting the value from 1. Doing so gives us 0.25. We can multiply this value by 100 to get the percent decrease in the odds of capturing a little brown bat. This can be achieved with this chunk of code in R.

```
oddsratio<-exp(-0.28354)#This gives us the odds ratio from the coefficient.

odds<-1-oddsratio#This gives us the odds.

percent_change<-odds*100#This gives us the percent change in the odds. We can use the sign
of the coefficient to interpret the direction of the change.
```

We know that it's a decrease because of the sign of our slope coefficient. This is not the same as a decrease in the probability of presence. If you would like more information on interpreting the coefficients, watch this helpful video: https://www.youtube.com/watch?v=Q1i58Ri6VsI&t=199s

To find the probability of presence for a given year, we can rearrange the logistic equation to solve for p. B_0 is the y intercept, B_1 is the slope coefficient, and x is the year.

$$p = \frac{\exp(B_0 + B_1 x)}{\exp(B_0 + B_1 x) + 1}$$

We can use the following code to solve for P.

```
odds2006<-exp(569.82594-0.2835*2006)#Plug in the y-intercept and subtract the year
coefficient multiplied by the year.
prob2006<-odds2006/(odds2006+1)
prob2006
```

Complete the R Lesson and Answer the Following Questions.

9. What was your favorite site name? I am partial to Bat Bug Bank.
10. By what percent did the odds decrease for 1 unit change in x for little brown bats? ~25%
11. What was the probability of presence in 2003 and 2019 for northern long-eared bats? Are the values the same as the focal paper? The logistic regression model predicts 0.878 in 2003 and 0.0717 in 2019. The values from the focal paper were 0.868 in 2003 to 0.262 in 2019.
12. Did the probability of presence of big brown bats increase or decrease? Is this similar to what the focal paper found? The probability of presence of big brown bats increased. The focal paper also noticed an increase in big brown bats.

Citations

Beck, J. M., & Morris, K. M. (2017). New County Records of Little Brown and Northern Long-eared Bats in Georgia. *Southeastern Naturalist*, *16*(2), N5–N8. https://www.jstor.org/stable/26454751

Ford, Clay. "Understanding Deviance Residuals | University of Virginia Library Research Data Services + Sciences." Data.library.virginia.edu, 22 Sept. 2022, data.library.virginia.edu/understanding-deviance-residuals/.

Hector, A. (2015). Generalized Linear Models for Data with Non-Normal Distributions. *Oxford University Press EBooks*, 121–140. https://doi.org/10.1093/acprof:oso/9780198729051.003.0009

Hector, A.(2021). The New Statistics with R: An Introduction for Biologists (2nd edn). (Oxford, 2021; online edn, Oxford Academic, 19 Aug. 2021), https://doi-org.proxy.library.vcu.edu/10.1093/oso/9780198798170.001.0001,

Johnson, C., Brown, D. J., Sanders, C., & Stihler, C. W. (2021). Long-term changes in occurrence, relative abundance, and reproductive fitness of bat species in relation to arrival of White-nose Syndrome in West Virginia, USA. *Ecology and Evolution*, *11*(18), 12453–12467. https://doi.org/10.1002/ece3.7991

Maness, C. (2016, March 28). Why does WNS affect American bats more than European bats? Retrieved April 18, 2023, from The Wildlife Society website: https://wildlife.org/why-does-wns-affect-american-bats-more-than-european-bats/#:~:text=They%20found%20that%20species%20like%20little%20brown%20bats

PSU. (2023). "Beyond Logistic Regression: Generalized Linear Models (GLM) | STAT 504." PennState: Statistics Online Courses, 2023, online.stat.psu.edu/stat504/lesson/beyond-logistic-regression-generalized-linear-models-glm#:~:text=Link%20Function%2C%20%CE%B7%20or%20g.

Shafi, A. (2021, May 18). What is a Generalised Additive Model? Retrieved from Medium website: https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a

Shively, R. (2017). Range And Roosting Ecology of the Little Brown Bat, Myotis lucifugus, In Interior and Northern Alaska. Retrieved from https://bioone.org/journals/northwestern-naturalist/volume-98/issue-2/NWN16-16.1/Range-And-Roosting-Ecology-of-the-Little-Brown-Bat-Myotis/10.1898/NWN16-16.1.short#:~:text=The%20Little%20Brown%20Bat%20is%20widely%20distributed%20across,%28Jones%20and%20others%202009%3B%20Boyles%20and%20others%202011%29.

TAYLOR, M. F. J., SUCKLING, K. F., & RACHLINSKI, J. J. (2005). The Effectiveness of the Endangered Species Act: A Quantitative Analysis. *BioScience*, *55*(4), 360. https://doi.org/10.1641/0006-3568(2005)055[0360:teotes]2.0.co;2

USFWS. (2023, April 17). State of the Bats | U.S. Fish & Wildlife Service. Retrieved from FWS.gov website: https://www.fws.gov/story/2023-04/state-bats

USGS. (2011). Beyond Billions: Threatened Bats are Worth Billions to Agriculture | U.S. Geological Survey. Retrieved April 22, 2023, from www.usgs.gov website: https://www.usgs.gov/media/audio/beyond-billions-threatened-bats-are-worth-billions-agriculture

Verant, M. L., Meteyer, C. U., Speakman, J. R., Cryan, P. M., Lorch, J. M., & Blehert, D. S. (2014). White-nose syndrome initiates a cascade of physiologic disturbances in the hibernating bat host. *BMC Physiology*, *14*. https://doi.org/10.1186/s12899-014-0010-4

White-Nose Syndrome Killed Over 90% of Three North American Bat Species | U.S. Geological Survey. (2021, April 21). Retrieved from www.usgs.gov website: https://www.usgs.gov/news/national-news-release/white-nose-syndrome-killed-over-90-three-north-american-bat-species