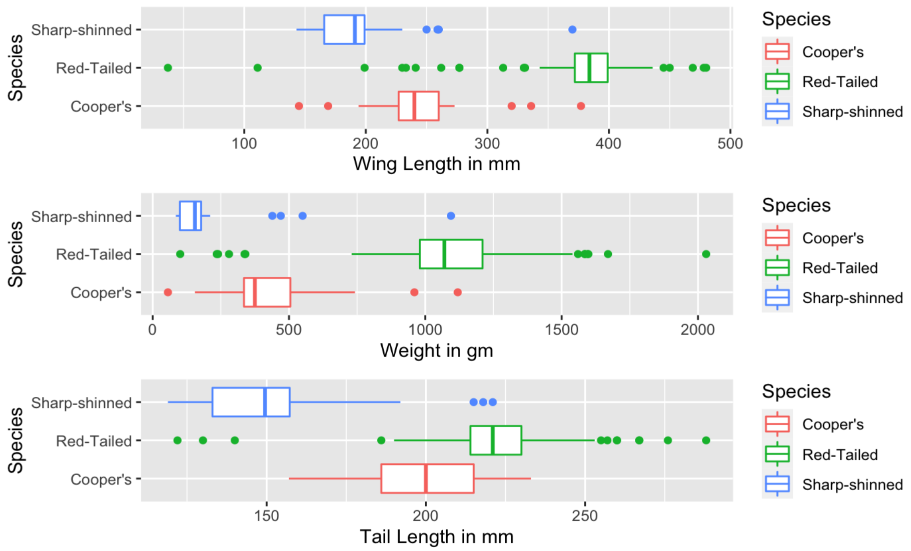
STA238 Final Project
Lucas Lyons

Hawks Data Analysis

In this project, I'm going to examine in detail the data set "Hawks" from the Stat2Data library, which is freely available on CRAN. By performing analysis on this data, I hope to answer the question "What are the trends and associations of hawk wing length, weight, and tail length for Red-tailed, Sharp-shinned, and Cooper's hawks?", based on the data collected. This data was collected from the years 1992-2002 by Cornell University students near LakeMacBride, Iowa. The data was collected under the supervision of professor Robert Black of Cornell, a professor of biology who studied the patterns of hawk migration among other things. The data set I'm using is a subset of a larger data set, featuring 908 observations from three species of hawk, the Red-tailed, Sharp-shinned, and Cooper's hawks. Other species of hawks with <10 observations were omitted from the data set. The set itself includes observations on the hawks' age, wing length, tail length, weight, and sex, as well as several other physical characteristics. All lengths are measured in millimetres, while weight is measured in grams. I will primarily focus on analysis of the variables wing length, weight, and tail length. Several of the other variables have many missing measurements and these variables are complete for most of the observations.

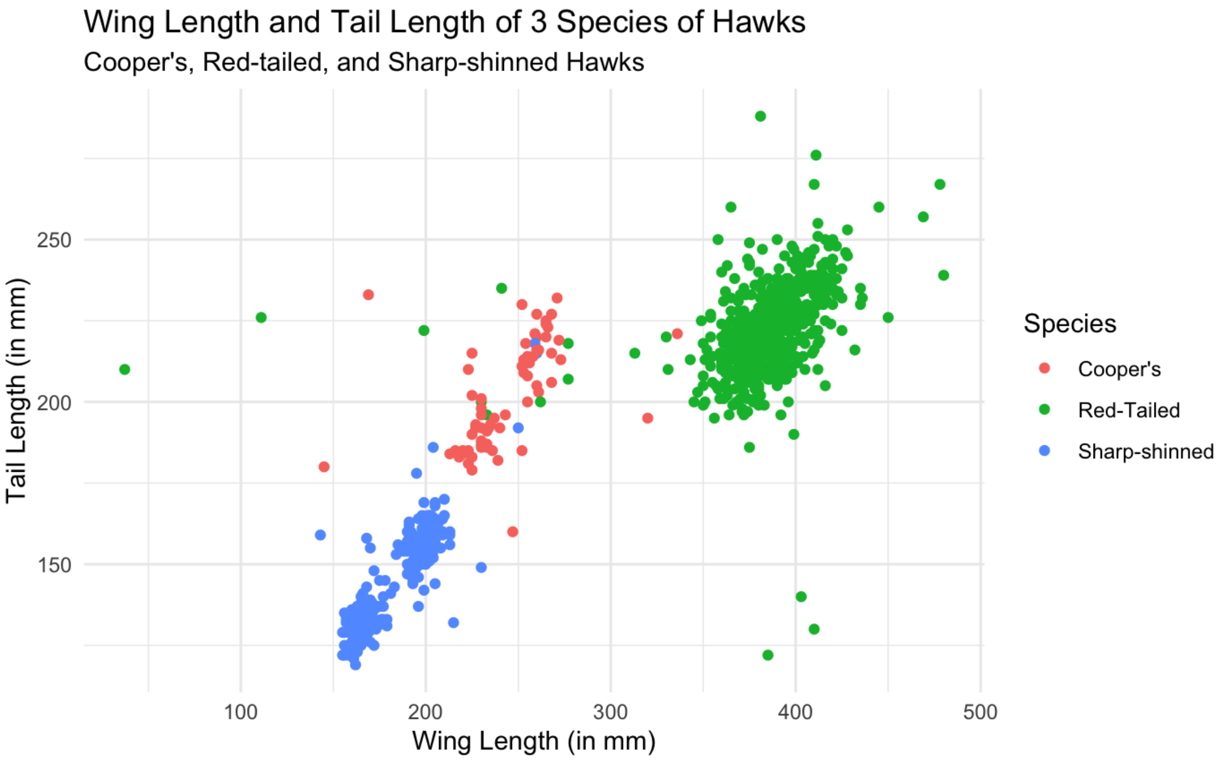
In order to answer this research question, I will first find confidence intervals for the population mean wing length, weight, and tail length for all three species of hawks. This will be helpful because it will enable us to make some comments as to the characteristics of the general hawk population based on our data sample. I will then perform a simple linear regression analysis on wing length and tail length of the three species to see if we can find what relation between the variables exists, if any.



A quick scan of our data indicated that 14 observations were missing data we required, so I removed these entries (Appendix 1). As a preliminary exploration, here you can see box plots of wing length, weight, and tail length for the three species of hawks.

From this data we can already see some trends. For example, we see that on average, Red-tailed hawks are the largest species while Sharp-shinned hawks are the smallest. There do appear to be many outliers for each variable and for all three species. Unfortunately, this may affect the quality of the inferences we are able to make from our data. Nevertheless, we'll do our best.

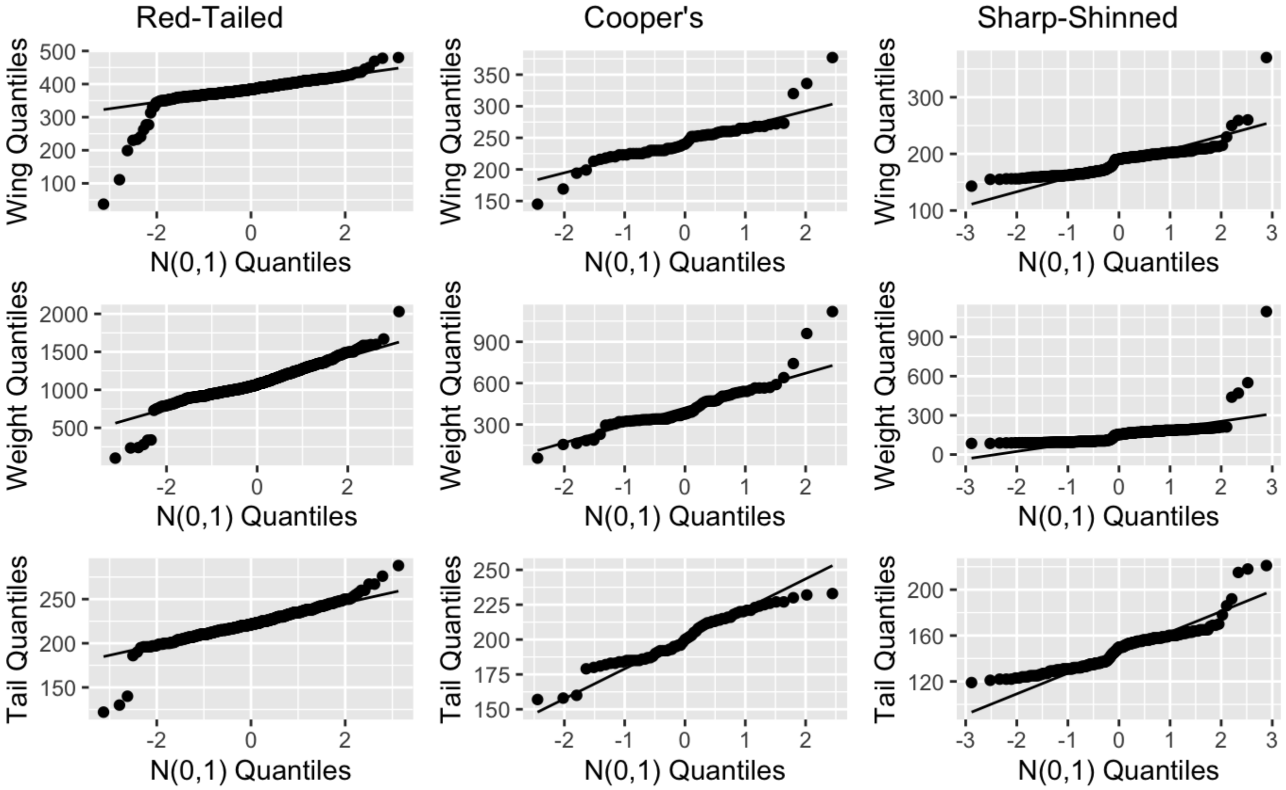
Let's compare wing length to tail length for the three species. I want to perform some simple linear regression on these two variables, so a quick glance can help us determine if this is at all feasible.



There does appear to be some type of linear relation between these two variables, so we're off to a promising start.

Before performing linear regression, I constructed 95% confidence intervals for the wing length, tail length, and weight of the three hawk species. In order to do so, it was necessary to verify whether or not the data was normally distributed. Had the data been normally distributed, the confidence intervals could have been computed using the standard normal distribution.

Unfortunately, the following quantile-quantile plots show that while some of the variables do appear nearly normally distributed, it seemed more prudent to bootstrap my data sample with B=2000 and use the studentized T-distribution to construct the confidence intervals (Appendix 2).

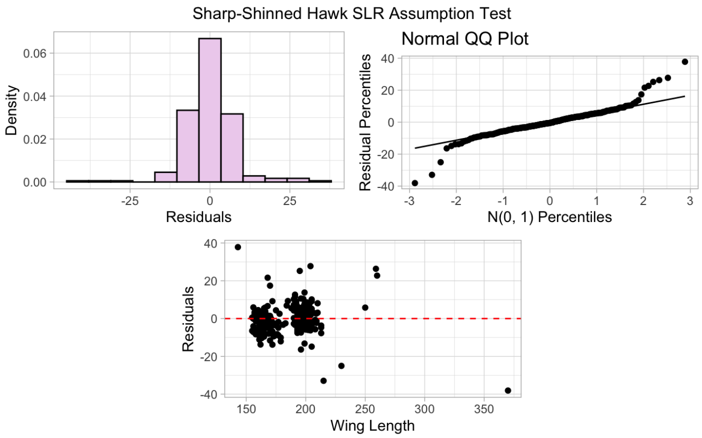


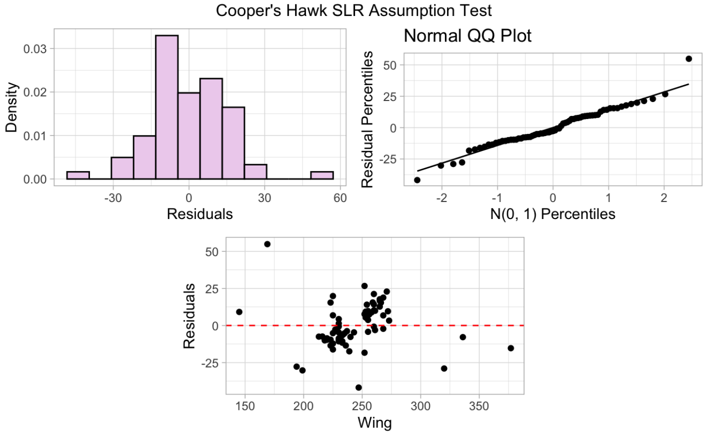
Here we can see that it tends to be the case that the tail ends of the samples do not conform to the normal distribution. This could be related to the numerous outliers present in the data, but since those outliers do not seem to be data entry errors, I decided to keep them in the data set.
See below a table of the computed confidence intervals.

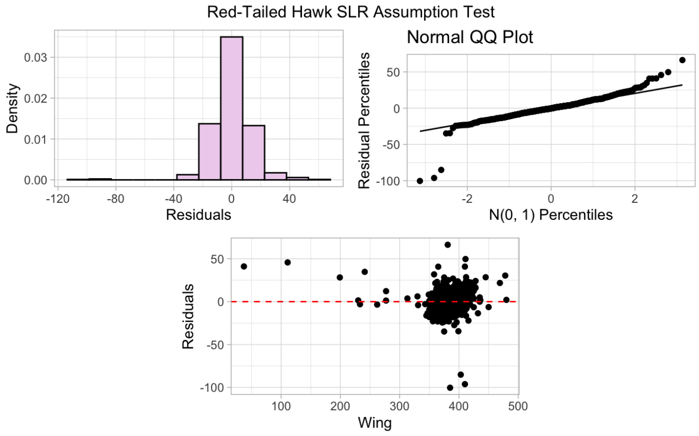
	Red-Tailed hawk	Cooper’s hawk	Sharp-shinned hawk
Wing length (mm)	(379.28, 386.45)	(234.21, 255.24)	(181.35, 188.90)
Weight (gm)	(1073.32, 1115.16)	(374.19, 479.83)	(138.34, 168.14)
Tail length (mm)	(220.56, 223.67)	(195.16, 206.53)	(144.24, 149.29)

There were a lot of samples in the data set for red-tailed and sharp-shinned hawks, so our confidence intervals are fairly narrow for these species. Cooper’s hawk has a smaller sample size, so the confidence intervals are larger. However, these estimates give us a helpful glimpse at the general characteristics of each population.

Now let’s perform some simple linear regression. For each species, we first determined whether it was appropriate to use SLR. The figure plotting wing length and tail length shows us that we have some reason to believe there is a linear relation between these two variables for each species. But recall that we need to verify three other assumptions in order to appropriately perform SLR: Independence of errors, normality of errors, and equal variance of errors. I used graphical methods to ensure these three assumptions were not violated.



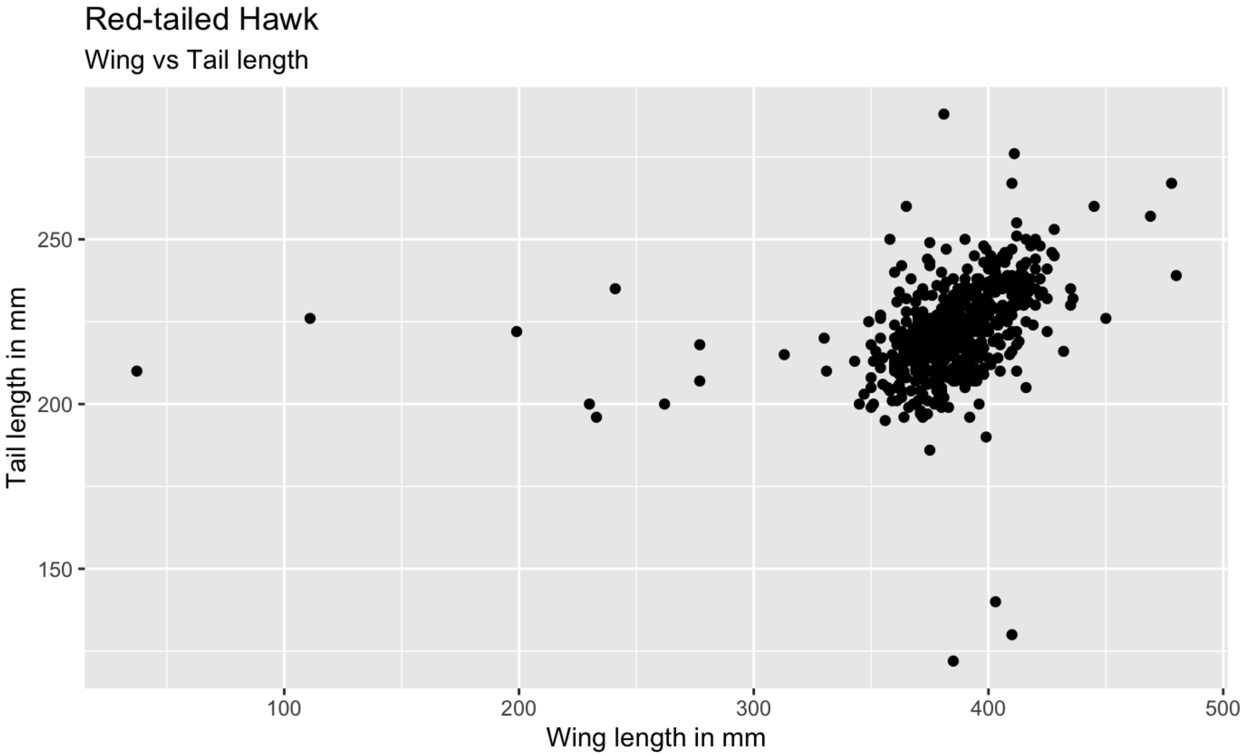




We use the histograms and the QQ plots to verify that the residuals are roughly normally distributed. In all 3 cases, it seems this condition is not violated. We then use wing length vs residual scatter plot to check if the residuals are independent and have equal variance. In other words, we want to see that the residuals have a random distribution about the line and that the variance of the residuals should be the same across the x axis. We see that the outliers create some difficulty in interpreting the data, but if we look at the main clusters of data they appear to not violate these assumptions. We therefore used R to create linear models with wing length as the predictor and tail length as the response variable. We could conclude the following 95% confidence intervals for our intercepts and slopes based on the data (Appendix 3):

	Red-tailed hawk	Sharp-shinned hawk	Cooper's hawk
Intercept	(149.30, 177.22)	(26.42, 42.22)	(98.65, 155.13)
Slope	(0.12, 0.19)	(0.56, 0.65)	(0.19, 0.42)

The sharp-shinned hawk measurements appear to give us the strongest correlation. Looking at the scatter plot for the red-tailed hawk, we might expect a stronger relation.



I hypothesize that the numerous outliers substantially affect the outcome of the SLR analysis. The large cluster of samples appears to have a relatively strong positive relation between wing and tail length but here we see that for red-tailed hawks the r value is lower than I expected. The confidence intervals for Cooper's hawks are wide because of the small sample size.

After performing some data analysis, it seems that wing length predicts tail length to some extent, but with varying degrees of accuracy depending on the species. Our sample for sharp-shinned hawks appeared to show the most robust relationship. This indicates that a larger sample might yield more informative results, because there may be fewer outliers in a larger data set and a relation could become more clear. But based on our data, we certainly can't make such a claim. We can claim that our calculated confidence intervals for our wing length, tail length, and weight are robust estimates for the population mean given our sample, in particular for sharp-shinned and red-tailed hawks as the sample size is high for those species. As I am learning, with real-world data sets, relations aren't always clear and easy to establish and even a relatively complete data set like this one can still provide challenges and setbacks to inferential analysis.