"Have the number of starlings and sparrows declined in rural areas from 1970 to 2008?"

"I confirm that the following report and associated code is my own work, except where clearly indicated."

1.Abstract

The dataset given is taken from a research study "Assessing factors associated with changes in the numbers of birds visiting gardens in winter: are predators partly to blame?" in which the abstract claims that there have been declines in many species of songbirds in recent years.[1] My research question is based on testing if this claim holds among house sparrows and starlings in urban/suburban areas. To do this I use the dataset provided, which shows the average count of house sparrows and starlings in volunteer's garden across the UK from 1970 to 2010, to investigate if these species of songbirds are declining.

It is important that scientists are aware if these species are on the decline and if so what the extent of the decline is because sparrows play a role in the functioning of ecosystems by maintaining the food web and an ecological balance. Feeding on seeds, grains, and larvae, the birds have proven to be an effective pest control agent. Also, starling are beneficial to the UK ecosystem as large flocks typical of this species can be beneficial to agriculture by controlling invertebrate pests.[2] Therefore, I have chosen to solely investigate the songbirds in rural areas as this is where they provide benefit to the UK ecosystem. Starlings are described as a nuisance through their noise and mess caused by their large urban roosts, so investigating the decline in urban areas has less significant ecological weighting as it has for rural areas.[3]

2.Data Wrangling

While there is no data to clean it is important that I put it into a format that is easy to deal with as I perform the parametric and non-parametric tests I want to perform. In this case as there are 6185 instances of data being recorded about the observation of birds so it is crucial to clean the data and manipulate it: we want the data to be interpretable, easy to understand in a visualisation.

In the data, the years are marked from 1 to 39 which is not very meaningful so I formatted them so they corresponded to the actual year (1970 to 2008). For the task, it is important to consider the species separately when assessing the decline of the birds over time to avoid Simpson's Paradox, as each species will have different population sizes, whereby both species may be appear to be declining over time but the overall trend of the bird observations may appear to increase which leads to an unmeaningful conclusion being drawn. Also, assessing both species separately allows me to draw more useful conclusions as to which has declined faster, if any. Consequently, I split the original data frame into two separate data frames (one for each species), as well as remove all the variables I am not concerned by such as temperature and northing & easting coordinates. I have created a new column in each data frame which takes the average number of that species observed per site in a given year, which forms the dependent variable that I will be using for this task as it makes the data easier to deal with rather than considering all 6185 sites at once (Figure 4, found later into the report plots these points).

3.Parametric Test

For my parametric test I am going to use a linear regression model to test the statistical significance of the independent variable "year". This is an appropriate parametric test to conduct in this case because it simply provides linear regression coefficients using the "lm" function in R. I am particularly interested in the slope coefficient that R outputs as this is a measurable way of representing an estimate in change in the average number of respective birds per site per year. If there is a decline in the average number of birds per site per year then we would expect our slope coefficient to be negative. Another benefit to using linear regression for this test is that the "summary" function in R conveniently displays the p-value associated with the coefficients so it would be very easy to say how confident I am in the value of the coefficient.

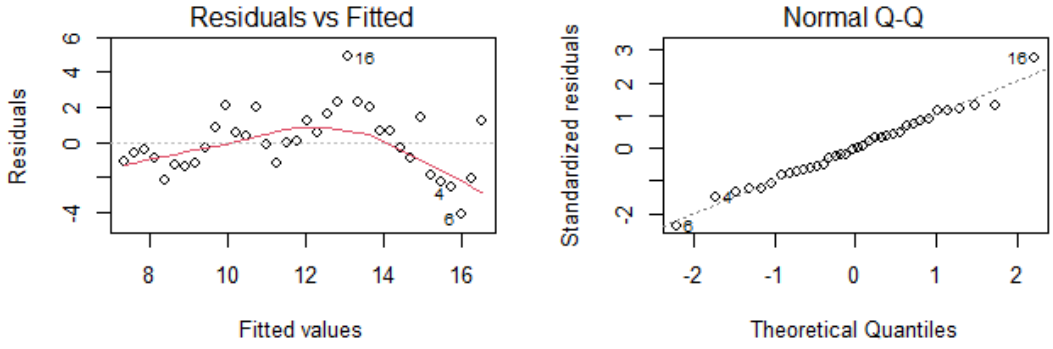

Figure 1a: shows two plots about house sparrows. Residual values vs fitted values & a normal quantile-quantile plot.

Method

Since linear regression assumes a linear relationship between the year (independent variable) and the count of respective species in volunteer's gardens per site per year (dependent variable), I have to firstly check that linearity is a reasonable assumption.

Firstly, I will comment on the reasonability of the linear assumption on the house sparrow data with the aid of Figure 1a. The residual values vs fitted values plot shows the lines of residuals tailing below the x-axis and the gradient of this line increasing in the negative direction suggesting there is heteroscedasticity. If I were to assume linearity then I would expect that the residuals would be well scattered below and above the x-axis which shows the variance is constant throughout all the years, and this is not the case here. Looking at the normal Q-Q plot we can see that the graph is reasonably linear suggesting that normality may not be an unreasonable assumption.

Now commenting on the validity of the linear assumption in starlings using Figure 1b, we can see that the residuals vs fitted plot follows very similarly to the correspondent plot for sparrow. It is clear there is some heteroscedasticity as the variance becomes greater at the tail ends of the data. However, normality is a reasonable assumption in this case as the normal Q-Q plot doesn't massively deviate from the line.

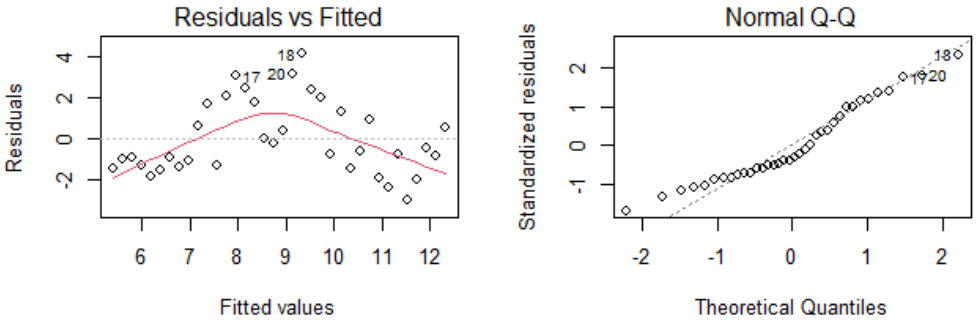

Figure 1b: shows two plots about starlings. Residual values vs fitted values, & a normal quantile-quantile plot.

Assuming linearity about both the starlings and sparrows observations over time seems very unreasonable. In order to deal with the heteroscedasticity I am going to perform a transformation on both sets of data to make the residuals exhibit a more constant variance across the whole range of fitted values to help me meet the assumptions of linear regression.

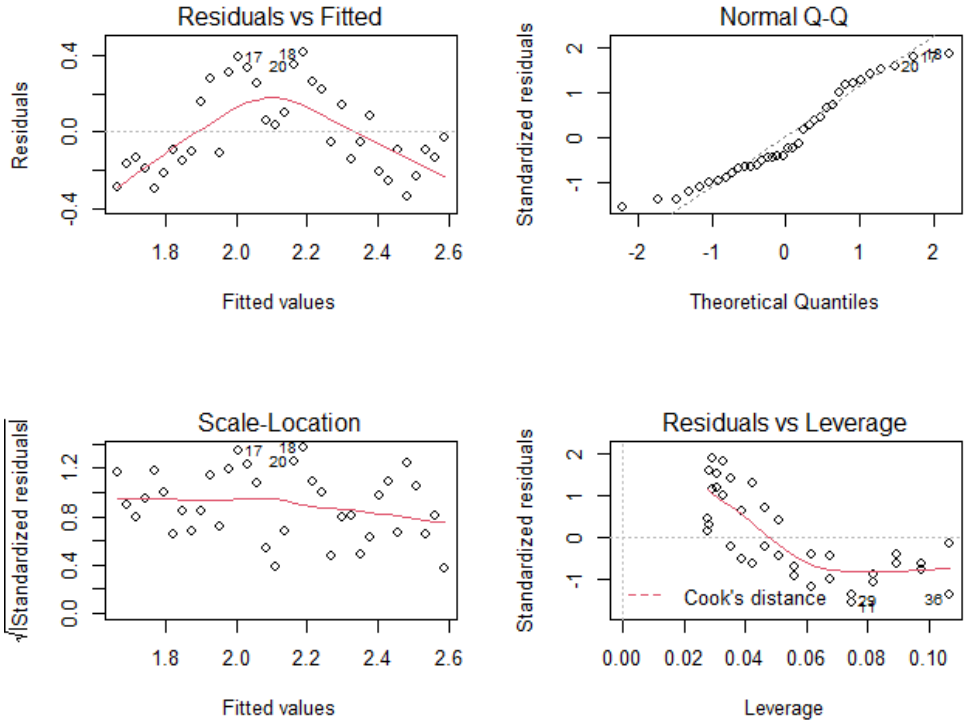

Figure 2: shows four plots on the sparrow data once we have transformed it by $\log_e$.

Figure 2 displays plots once the sparrow data has been transformed via a logarithm with base *e*. I can see that the normality of the data has remained very similar and therefore it is still reasonable to assume normality for the data. The most pressing issue I had before the transformation was the heteroscedasticity evident in the residuals vs fitted values plot. The line of residuals on the residuals vs fitted values plot follows the line along zero much more closely than in Figure 1a. The vertical spread of the residuals vs fitted values has significantly decreased after applying a log transformation as shown in the scale of the y-axis of the graph meaning variance is much more constant. This is also

shown in the scale-location plot as the line of best fit for the square root of standard residuals appears to be horizontal so there is no evidence of a trend. As a result, I think that constant variance is a reasonable assumption for the log-transformed data. I have omitted the similar plots for the transformed starling data (they can be found in the R code), but it follows very similarly to the transformed sparrow data and the same conclusions can be drawn.

The last assumption I have to make is whether it is reasonable to assume independence among the observation of birds per site in a given year. Given that "gardens are chosen carefully from existing members of the Garden BirdWatch survey ensuring good coverage across the UK", I think that it is a reasonable to assume independence of the observations.[4] Therefore, I have declared all the assumptions I must to apply a linear regression to the transformed data.

I am testing the value of the slope coefficient. The model I form will be $\log_e(y) = a + b * x$ where y is the average number of birds per site, a is the intercept, b is the slope coefficient and x is the independent variable year on the domain x = 1970, 1971, …, 2008. Back-transforming gives the model $y = \exp(a + b * x)$. If $b<0$, the average number of birds per rural site has decreased over time and the more negative b is, the greater the rate of decrease is. Therefore, the hypotheses for the parametric test are H_0: $b = 0$ & H_1: $b < 0$.

The pseudocode for this test is as follows:

1. Produce plots to check linearity is reasonable.
2. Log transform the data.
3. Create a linear model for the log-transformed data.
4. Check the slope coefficient is negative and the p-value is significant.

Results

Applying a linear model to the transformed sparrow data, the "year" coefficient (b) is -0.02432407 with a p-value of < 3.7656e-11 and for the transformed starling data the slope coefficient is -0.02629017 with a p-value of < 1.345507e-08. These extremely small p-values indicate that there has been a statistically significant decline in the average number of birds per site over the years so we can reject H_0 at the 1% significance level. As a result, the negative slope coefficient is meaningful to the model of the average numbers of each bird per site over time

4.Non-Parametric Test

I am using non-parametric bootstrapping as a non-parametric test to see if there is any statistical evidence that the number of starlings and sparrows have declined over time in rural areas. The bootstrap works by resampling the data 1000 times with the same number of datapoints per resample as the original data, but allows replacement so that datapoints are more than likely to be replicated. I then apply a linear model to each set of resampled data to obtain the slope coefficient. As shown in the previous section, I cannot assume linearity for the original data so I cannot assume linearity for the resampled data. Therefore, I perform the bootstrap on the log-transformed data . For each set of resampled data I apply a linear model and store the slope coefficients into a vector and deduce the critical values for a 99% confidence interval.

Method

Outlined below is the pseudocode used for running the non-parametric bootstrap on the sparrow data set.

1. Initialise a vector with length 1000.
2. Create a random sample with replacement of the indexes in the sparrow data frame.
3. Create two vectors using the resampled indexes with the corresponding year and log of the average number of species per site for that year.
4. Run a linear regression model on the resampled data on the resampled years.
5. Store the value of the slope coefficient to the first value of the initial vector.
6. Repeat steps 2-5 another 999 times storing the i-th repeat to the i-th index of the initial vector.
7. Take the 0.05% and 99.5% quantiles of the simulated distribution.
8. Plot a histogram of the slope coefficient values from the 1000 samples.
9. Plot onto the histogram the confidence intervals calculated in step 7.

I have created a function in my code, 'lm_bootstrap', which performs the bootstrap with arguments 'nboot' (number of times I want to perform bootstrap) and 'df' (the data frame I want to use). The data frame must be in the same

format as the way I have formatted the starling and sparrow datasets, i.e. the year of the sample in the first column and the dependent variable of the linear regression model in the fifth column.

Results

For the sparrows, the critical values are -0.03216758 and -0.01665644, and for the starlings they are -0.03491314 and -0.01751297. As a result, non-parametric bootstrap provides the same result as my parametric test that there is very strong statistical evidence (99% confidence interval) to suggest the number of both sparrow and starling have declined since 1970.

For the sparrows, the critical values are -0.03216758 and -0.01665644, and for the starlings they are -0.03491314 and -0.01751297. As a result, non-parametric bootstrap provides the same result as my parametric test that there is very strong statistical evidence (99% confidence interval) to suggest the number of both sparrow and starling have declined since 1970.

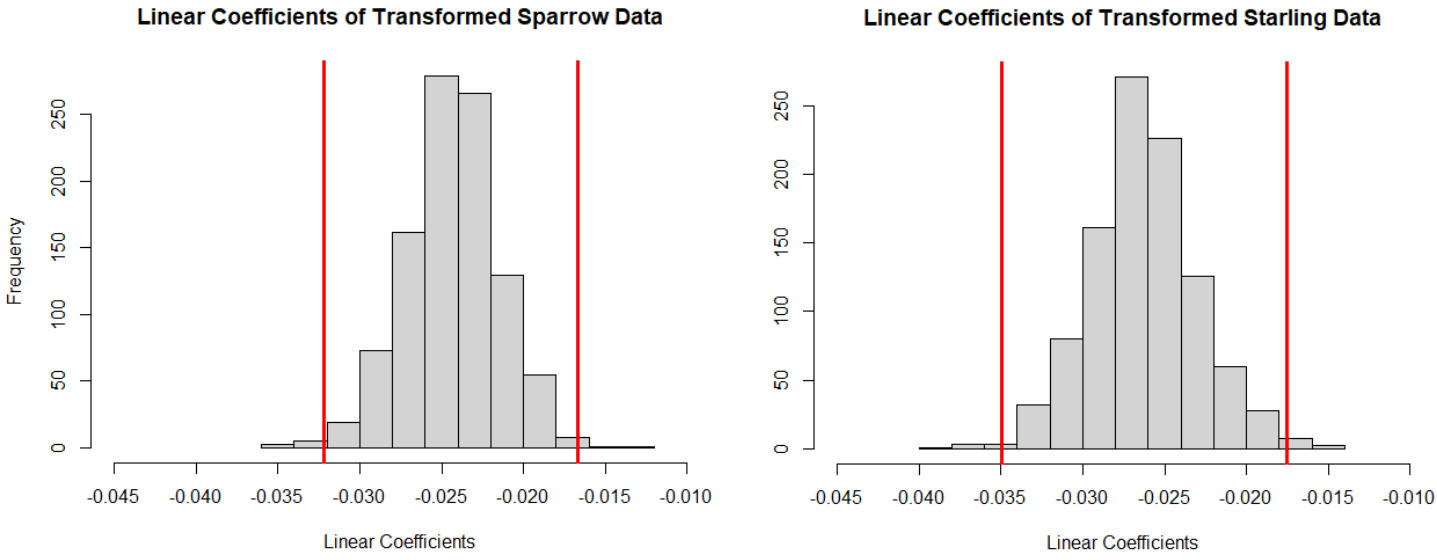

Figure 3

The histograms in Figure 3 show the sets of slope coefficients for the 1000 linear models in each case. There are no observed coefficients greater than approximately -0.012 < 0 for sparrows and this follows similarly for the starling data. For both sets of data it is very clear that there is very strong evidence for a decline in the species over the years since there are no linear models with slope coefficients equal to or greater than 0.

5.Conclusion

Since both the linear regression parametric test and the non-parametric test provided the same result of there being very strong evidence of a decline in the number of starlings and sparrows from 1970 to 2008. The p-values for both tests were significantly small so there is very clearly a negative trend in the number of these songbirds in the 40 year frame the data was recorded.

Figure 4 displays the average number of birds per site along with the parametric log-transformed linear model. It is clear that both starlings and sparrows have decreased over time at a very similar rate. While this decline could be for any number of reasons such as a reduction in the prey of the birds or increased deforestation in rural areas, it is important that we have shown there is a decline. This shows for conservationists, that measures to protect each species are of equal importance.

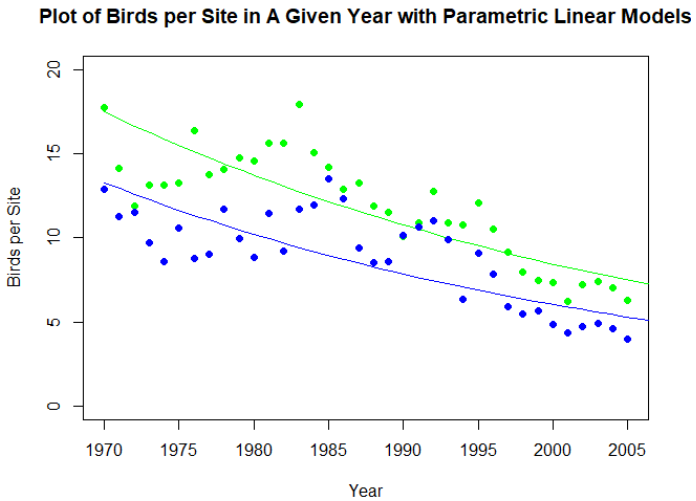

Figure 4: Sparrow represented by green and starling by blue

Therefore, parties concerned with the decline of these birds must work together to prevent this decline as if these birds go extinct it could have a vast effect on the ecosystem throughout Britain as well as on crops and profits of farmers as they are losing their natural "pest control agent".

6.References

[1]**Drayad**
https://datadryad.org/stash/dataset/doi:10.5061/dryad.v8j1144 **(accessed 16th October 2023)**

[2] **Wildlife SOS**
https://wildlifesos.org/animals/house-sparrow-the-ultimate-urban-dweller/#:~:text=Moreover%2C%20they%20generally%20occupy%20buildings,an%20effective%20pest%20control%20agent **(accessed 16th October 2023)**

[3]**Wikipedia**
https://en.wikipedia.org/wiki/Common_starling#:~:text=Large%20flocks%20typical%20of%20this,by%20their%20large%20urban%20roosts **(accessed 16th October 2023)**

[4]**British Trust for Ornithology**
https://www.bto.org/our-science/projects/gbfs **(accessed 17th October 2023)**