Analysing and Predicting the Size of Blackbirds using EDA and Linear Regression

Author: 35315878

29/10/2019

1 Abstract

It has been shown that a reasonable estimate of the size of a blackbird can be obtained from measurements of its wing, given that the weight of a blackbird varies during the course of the year. In this paper, we begin by utilising a common data mining technique known as Exploratory Data Analysis (EDA) to analyse a data set of blackbirds collected in a garden over a period of 25 years. This involves understanding the types of variables in the data, calculating summary statistics and using visual methods to draw further insights from the data. Following EDA, we quantify the effect of the variables on the size of a blackbird using a quantitative method known as linear regression, which enables us to discover linear relationships between dependent and independent variables. We investigate how these variables affect the size of a given blackbird using the fitted linear model and highlight ones with the most significance. We will use R, a statistical analysis package to conduct our EDA and implement the linear regression model. It was found that the variables which influence the size of a blackbird were the age, weight and sex of the blackbird, with the time of year also influencing its size.

2 Introduction

Table 1 shows the variables present in the blackbirds data set:

Variable	Description
Ring.number	the unique bird identifier
Sex	male (M), female (F) or unknown (U)
Age	juvenile (J), first-year (F), adult (A) or unknown (U)
Wing	length of wing (in millimetres)
Weight	weight of bird (in grams)
Day	the day the bird was recorded
Month	the month the bird was recorded
Year	the year the bird was recorded
Time	in GMT (winter months) and BST (summer months)

Table 1: Blackbird data set variable descriptions.

After analysing the types of each variable we have 3 categorical variables: `Ring.number`, `Sex` and `Age`. R interprets these as factors as they take on a limited set of possible values. `Ring.number` contains 2141 levels (i.e. 2141 bird identifiers). `Age` takes on values J, F, A and U as explained in Table 1. `Sex` takes on values M, F or U. The rest of the variables are all numerical integer values.

The key question to answer in this paper is which of the variables outlined in Table 1 affect the size of a blackbird and by how much? A linear regression model will be implemented to answer this question, by quantifying the size of the effect of each of the variables in the data set on the size of a blackbird.

3 Methods

3.1 Exploratory Data Analysis

Analysing the blackbirds data loaded into R we obtain a sample size of 4123 observations described by 9 distinct variables. It must be noted that measurements taken on a specific day for a given bird may be repeated in the data, hence the number of independent observations will be less (although for our analyses we ignore this).

The R function `summary()` was used on the blackbirds data to yield the following summary statistics:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Wing	3,872	129.712	4.270	114.0	127.0	133.0	144.0
Weight	4,083	107.063	12.639	60.0	98.0	116.0	152.0
Day	4,123	16.049	9.080	1	8	24	31
Month	4,123	4.411	3.775	1	1	7	12
Year	4,123	2,004.163	7.705	1,988	1,997	2,010	2,015
Time	4,123	11.494	3.214	0	9	14	21

Table 2: Summary statistics of blackbirds data

From Table 2 we can see some of the missing values are present in the Wing and Weight variables. Using the *is.na*() function on our blackbirds data frame calculated 0.7% of the data contained missing values.

We can analyse how the categorical variables in the data affect the wing length of a blackbird. The following box plot shows how the age of a blackbird affects the length of its wing:

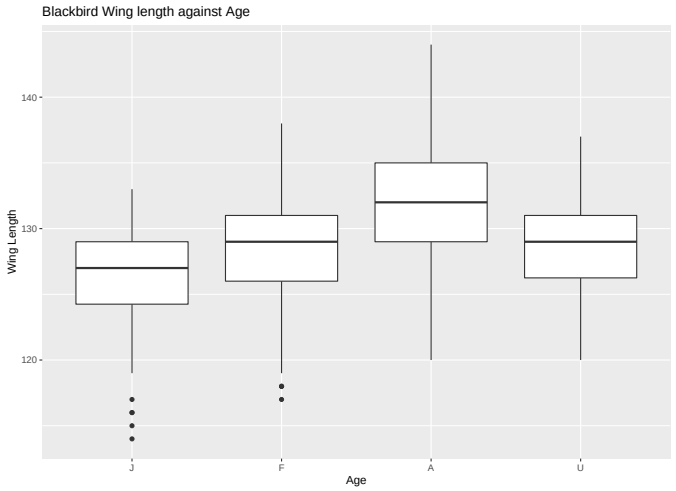

Figure 1: Box plot of blackbird age and wing length

The box plot in Figure 1 shows how the wing length changes as the blackbird ages. We notice a general increase in blackbird wing length as the blackbird ages from juvenile to adult where the largest wing lengths are found in adults and the smallest lengths are found in juvenile blackbirds.

We can create another box plot to analyse how the sex of a blackbird affects the length of its wing:

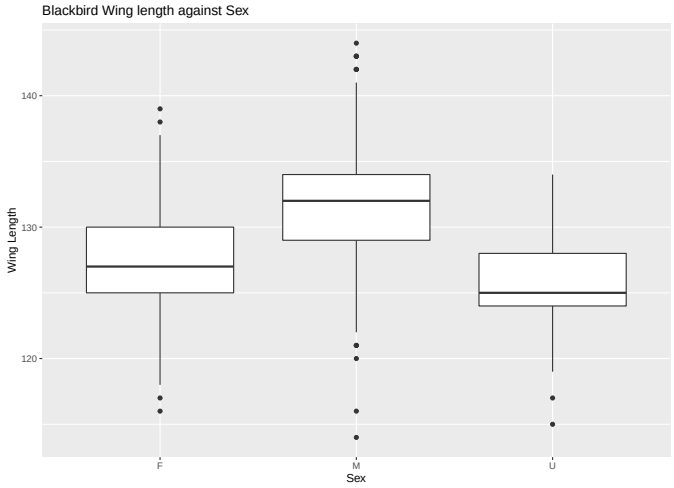

Figure 2: Box plot of blackbird sex and wing length

The box plot in Figure 2 shows how the wing length changes amongst blackbirds of different sexes. We notice that male blackbirds generally have a larger wing length compared to female blackbirds.

Further analysis showed birds of the same identifier were recorded multiple times - this may have an effect on the final regression model due to repeated measurements.

3.2 Linear Regression

The linear regression model is used to model the relationship between a dependent (response) variable and one or more independent (explanatory) variables. The model is based on the linear equation $y = mx + c$, where m represents the gradient of the line, c represents the y-intercept and x and y represent the independent and dependent variables respectively.

The following equation outlines the simple linear regression model with 1 explanatory variable x_i for the response variable y_i:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where β_0 and β_1 are regression coefficients (which will be estimated from the sample data) and ϵ_i are independent and identically distributed random variables with $\epsilon_i \sim \text{Normal}(0, \sigma^2)$.

To determine a best fit line through our data, we must minimise the sum of squared residuals (also known as residual sum of squares). The residuals represent the differences between the observed response variable values and the fitted line. To find estimates of the parameters β_0 and β_1 we minimise the sum of squared errors using the following equation:

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_i \epsilon_i^2$$

We will utilise the R function `lm()` to facilitate the implementation of our linear model which implements the previously described techniques to fit a linear model to our data.

4 Results and Discussion

Calling the `summary()` function on our regression model yields the following results:

Explanatory Variable	Estimate	Std. Error	t	$Pr(>\|t\|)$	Significance
(Intercept)	61.117507	12.490069	4.893	1.03e-06	***
Male	2.729596	0.309896	8.808	$<$ 2e-16	***
Female	-1.617059	0.317052	-5.100	3.56e-07	***
Juvenile	-2.042432	0.384311	-5.315	1.13e-07	***
FirstYear	-1.864786	0.343492	-5.429	6.02e-08	***
Adult	1.344225	0.343634	3.912	9.32e-05	***
Weight	0.104083	0.004118	25.274	$<$ 2e-16	***
Day	0.008947	0.005365	1.668	0.09548	.
Month	-0.066121	0.013203	-5.008	5.75e-07	***
Year	0.028730	0.006237	4.606	4.23e-06	***
Time	-0.040562	0.014918	-2.719	0.00658	**

Table 3: Linear Regression Results

From the results in Table 3 we can see almost all variables are highly significant due to the three asterisks, that is, we observe very small p values ($Pr(>|t|)$). These small values mean that is is unlikely we will observe a relationship between a given explanatory variable and the response variable (Wing) due to chance. The variables Male and and Weight were found to have very small p-values ($<$ 2e-16) meaning these represent good predictors of a bird's wing length. Conversely for the Day variable, we observe a larger p-value larger than our significance level of 0.05 meaning it is less significant than the other variables. This means the Day isn't a very good predictor of a given bird's wing length.