**Juan Villegas**

**Project 1A**

1. Run the following command: summary(bird). How many missing values (NA) are identified for each variable in the bird dataframe?

Wing Weight

Min. :65.00 Min. :19.50

1st Qu.:69.00 1st Qu.:22.90

Median :71.00 Median :24.15

Mean :70.89 Mean :24.34

3rd Qu.:73.00 3rd Qu.:25.73

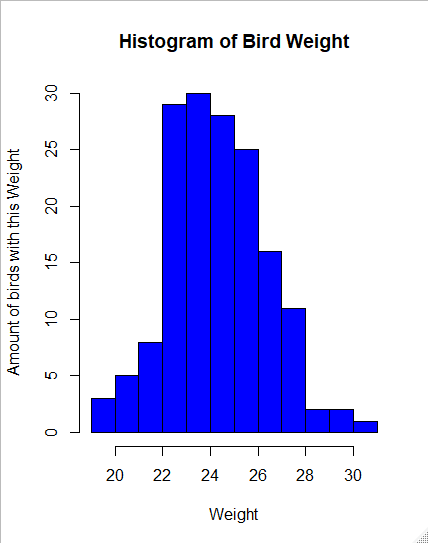
Max. :77.00 Max. :30.60

NA's :2 NA's :1

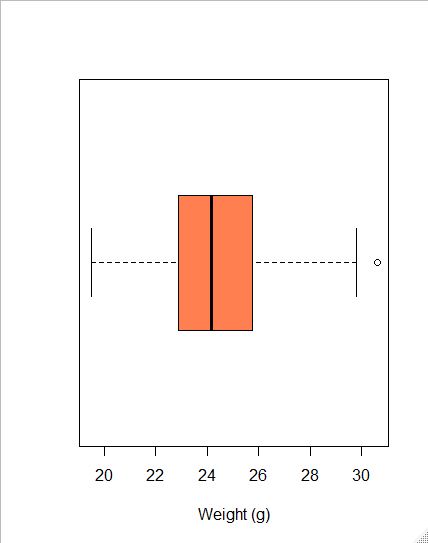
1 for weight, 2 for wing.

1. Is it reasonable to assume that bird weights are normally distributed? (We need this to be true for inference.) Answer this question after completing parts a – c.

The birds’ weights are very close to being normally distributed. Our boxplot shows that we do have an outlier which skews our data slightly to the right.

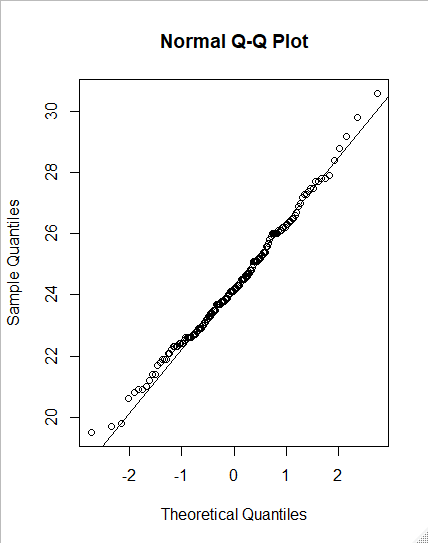
1. Make a histogram of weight. Use the xlab command to label the horizontal axis and the col command to give the graphic display an interesting color. Describe the shape of the weight data.

This graph is a unimodal and normal.

b. Are there any outliers? Make a boxplot. I prefer horizontal boxplots (but you might prefer them to be vertical). Here’s a sample command:

boxplot(bird$Weight, horizontal = TRUE, xlab = “Weight (g), col = “coral”)

There is an outlier. Looks like a bird weighing around 32 grams



c. Make a normal quantile plot for Weight:

qqnorm(bird$Weight)

qqline(bird$Weight)

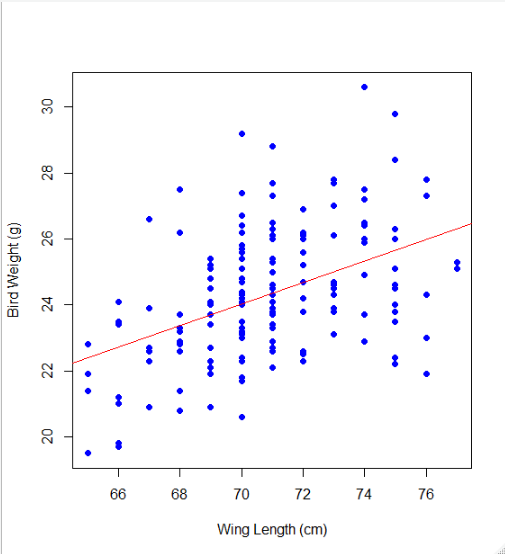
3. a. Make a scatterplot of Weight versus Wing length. Use solid dots (pch = 16) and choose a color (col = “pick color”). Label the horizontal axis “Wing length (cm)” and the vertical axis “Weight (cm).” (No need to add the scatterplot here. You’ll do that in part (c).)

If your scatterplot looks a bit strange, that’s because wing lengths have been rounded to the nearest half centimeter.

b. Fit a least-squares regression line to the data in your scatterplot. (Save the results of the lm model as Model1.) What is its equation? Interpret the value of the slope in the context of these data.

Y = 1.0728 + 0.3279x

Birds that do not have any wing length, weigh approx 1.0728g. For every cm that a bird’s wing length is, the weight of the bird in g, increases by 0.3279. For example, if a bird’s wing length is 70 cm, the bird weighs approx., 24.0258g (1.0728 + 0.3279(70)).

c. Overlay a plot of the least-squares regression line on your scatterplot from (a). Insert it here.

4. To check the adequacy of a linear model, make a residual plot (residuals versus explanatory variable which is equivalent to residuals versus fitted values). If the dots in this plot appear randomly scattered with no strong patterns (about half of the residuals should be positive and half negative), then the model is adequate to describe the pattern in the data.

a. Make a residual plot: a scatterplot of the residuals versus Wing. What problem do you encounter? Go back to the summary command that you ran at the beginning of this project. Why do you think you got this message?

Chart, scatter chart

Description automatically generated**x & y length differs.** We got this because there are 2 NA values in the ‘Wings’   
Category.

b. Instead, plot the residuals versus the fitted values. Add a horizontal line at y = 0. The command to add this line is: abline (0,0). Based on this residual plot, is the least-squares regression model adequate to describe the pattern in these data?

I believe that the Least Squares Regression Model is adequate to describe the pattern in these data because it looks like the data is almost evenly distributed between the positive and negative residuals.

c. Run the anova command on your regression model (the output from the lm command). Determine SSR, SSE, and SST. Based on these values calculate the value of  and interpret the results in the context of the bird data.

Text, letter

Description automatically generated

SSR = 124.49

SSE = 511.70

SST = 124.49 + 511.70 = 636.19

 = 124.49/636.19 = .195668054

19.57% of the variability in the bird weight (g) can be explained with the least square line on the bird wing length (cm).

d. Run the summary command on your regression model. Determine MSE and ?

Text

Description automatically generated

MSE = SSE/df

MSE = **511.70**/157 = 3.25923567

 = sqrt(MSE)

Sqrt(3.25923567) = 1.80533533

5. Let’s return to analyzing one variable at a time.

1. Calculate the mean and standard deviation of Weight (e.g., mean(bird$Weight) and sd(bird$Weight). What problem did you encounter?

There is N/A values

b. Run the command help(mean). What you will find is that the default in R is na.rm = FALSE. Which means R tries to sum the Weight data with NA (a missing value) as part of that sum. So, we will fix that problem as shown below. You can do the same for the sd command.

mean(bird$Weight, na.rm = TRUE)

> mean(bird$Weight, na.rm = TRUE)

[1] 24.3375

> sd(bird$Weight, na.rm = TRUE)

[1] 2.020267

> mean(bird$Wing, na.rm = TRUE)

[1] 70.88679

> sd(bird$Wing, na.rm = TRUE)

[1] 2.70719

1. Calculate the correlation between Weight and Wing length: cor(bird$Weight, bird$Weight). What problem did you encounter? Here’s the fix:

Also, N/A Values

cor(bird$Weight,bird$Wing,use = “complete.obs”)

> cor(bird$Weight,bird$Wing, use = "complete.obs")

[1] 0.4423538

Up to this point, we have encountered several problems with running commands where there are missing values in the data. There are several ways to deal with this problem. One method is to create a new dataframe that only includes complete cases. (This may not be the best approach because you may lose too much information from your data. However, in this case, there are only two variables and a small number of missing values.) In question 6, we’ll create the dataframe modbird (to indicate we have modified the original dataframe).

6. a. First, we create a variable that shows which records have both weight and wing data values.

Comp <- complete.cases(bird)

Summary(Comp)

Mode FALSE TRUE

logical 2 159

Explain what this command has done.

It shows us the number of values in our data that are N/A… or false (2)

b. Next, we create two variables weight and wing, which will lie outside of the bird dataframe.

weight <- bird$Weight[Comp ==TRUE] (the double =’s is not a mistake!)

wing <- bird$Wing[Comp ==TRUE]

modbird <- data.frame(wing, weight)

summary(modbird)

wing weight

Min. :65.00 Min. :19.50

1st Qu.:69.00 1st Qu.:22.90

Median :71.00 Median :24.10

Mean :70.89 Mean :24.32

3rd Qu.:73.00 3rd Qu.:25.65

Max. :77.00 Max. :30.60

How does the modbird dataframe compare to the bird dataframe?

The Modbird dataframe took out the N/A values in our old dataframe.

7. a. Using the modbird dataframe, fit a least-squares regression line to the data. Save the results of the lm command as Model2. Compare the equation to the one that you got from question 3(b).   
 y = 1.0728 + 0.3279x

I got the same equation as 3b.

1. Make a scatterplot of the residuals versus the explanatory variable. (You were unable to do this in question 4(a).) Compare this plot to the residual plot from question 4(b).

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

The only difference that I can see is that the X values have been increased by 50 & the x name is different. (Wing sz vs Fitted values).

c. You can get R to create a variety of residual plots. There are 4 plots. We’ll produce them in a 2 by 2 display. Here’s how:

par(mfrow = c(2,2))

plot(Model2)

par(mfrow = c(1,1)) Don’t forget this step!

Diagram, schematic

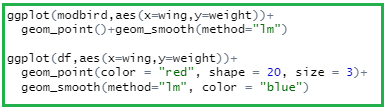
Description automatically generated

8. To end Part A of Project 1, let’s look at a plot using ggplot2. This will require that you download a package.

install.packages(“ggplot2”)

library(“ggplot2”)

Now that ggplot2 has been loaded, we will make a scatterplot of the data, overlay the regression equation, and add a confidence band around the least-squares line. In the second set of code, I have added my choice of colors.



**Comment:** Start by entering the name of the dataframe, then define the x and y variables, specify the type of graph (we are plotting points), and then the smoothing method used to produce the line.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated