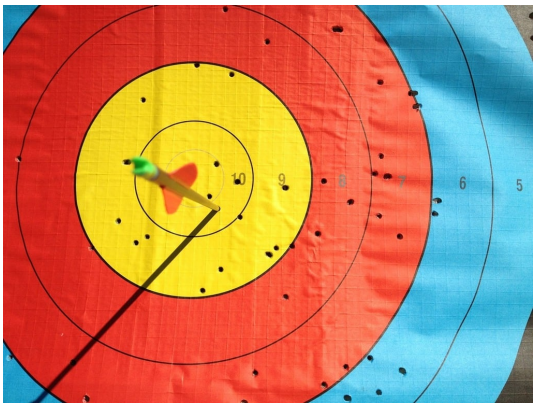
MATHS 7107 Data Taming
Assignment 2

Trimester 2 2024



1 Background

A local archery club has come to see their friendly, neighbourhood data scientist (ie. you). They want some help to analyse some data that they've collected, in exchange for paying you a lot of money.

They are trying to recruit new members, and one of the main questions they are asked by potential recruits is: **"how long does it take to become a good archer?"** The club doesn't have a good answer to this question yet, but they're hoping you can try to come up with some quantitative statements using the data from a recent tournament.

On the 15th June the club held a tournament for all their members. The tournament started at 7pm, giving the competitors time to have some food and a rest after the scheduled training session that same day. The club has three separate venues, and at each venue 555 archers competed in the tournament, where each archer shot up to 70 arrows. The club recorded the archers' names, the date they started practising archery and their performance at the tournament. Importantly, not all the club members have given their permission for their data to be used, so you need to make sure you de-identify the data before you get started.

Using this data, try to provide a good answer to the club's question. Conveniently for you, they have just started using `R` and `R Markdown`, so they want your report as a PDF generated using `R Markdown`. The club's president also studied Data Taming last trimester, so she wants you to only use commands from the course, so that she can easily see what analysis you've done. In your `R Markdown` code chunks: make sure that you **do not** set `echo = FALSE` so that she can see what `R` code you used to generate your output. But of course, she doesn't want to see irrelevant warnings or messages.

1.1 Number of digits

When writing your own text, or **USING** the output from `R`:

- For integer results, report the whole integer.
- For non-integers with absolute value > 1: use 2 decimal places
- For non-integers with absolute value < 1: use 3 significant figures.

  For example:

- $135.5681 \approx 135.57$
- $-0.0004586 \approx -0.000459$

Exceptions:

- If you're just **PRINTING** the output from `R`, then just keep the output as it is.
  - But if you have `R` do the rounding for you then you need to conform to these two conventions listed above.
- If your data has fewer digits of precision than specified above (eg. because of the way it was stored in the original data, or because of the way it was calculated) then only report that level of precision.

2 The data

The club has three datasets labelled `archery_0.csv`, `archery_1.csv` and `archery_2.csv` (one from each of the venues). Each dataset contains 3 columns:

- `Archer`: the name of the archer
- `Commenced`: the date the archer started practising archery.
- `RES`: the archer's tournament result, stating how many arrows they shot and the number of times they hit the target.

Each dataset has data on 555 archers. The club tried hard to ensure that the data was all entered correctly, but if you find any data that is erroneous, then remove the entire row before you do any analysis. Make sure you explain why you remove any data.

> **IMPORTANT!**
>
> Make sure you only remove data that you must remove. Do not just delete data because it is inconvenient. You must have specific instructions from the client, or it must be an impossible value, before you remove any data from your analysis. Even then, you need to describe why it was removed.

3 Your job

The club's question of **"how long does it take to become a good archer?"** is not well-posed for data analysis. But let's start trying to help the client as well as we can. We'll try to make some predictions for how good someone can get after some years of training, and we'll also provide a 90% confidence interval for our predictions.

> **Note**
>
> Make sure you write text to explain what you are doing at each point and why you are doing it. You need to justify all the things you do or claim. Also describe the results.

1. Load the correct dataset and save it as a tibble. Output the first 10 lines of the dataset and the dimensions of the data set.
2. Before we do anything else, let's tidy up the data a bit:
   (a) De-identify the data, by replacing the `Archer` column with another column called `ID`, which contains a unique number from 1 to 555. Be careful not to change the order of the data, ie. make sure that the person in row number 1 gets the number "1" and the person in row 2 gets the number "2", etc...
   (b) Use the `lubridate` package to create a new column (just to the right of the existing date column) called `XP`, that gives the number of days of archery experience that the archer had on the day of the tournament. (Make sure you think about whether to include the end points.)

(c) Replace the `RES` column with 2 new columns at the right of the dataset: `Arrows` containing the number of shots, and `Targets` containing the number of hits. (Put `Targets` as the far right column.)

When you've done these steps output the first 10 lines of the dataset and the dimensions of the data set.

3. Using dot points, identify what types of variables we now have in our data set, i.e., "Quantitative Discrete", "Quantitative Continuous", "Categorical Nominal", "Categorical Ordinal". (Don't just describe what data type they are in the data set — you need to think about the type of variable in the context of the meaning of the data.) Make sure you provide some justification for your choice of variable types.

   - Don't just provide vague statements, but be very concrete about describing this particular set of data.

4. Now it's time to clean and tame our data.

   - Make your data set conform to the Tame Data conventions on page 3 of Module 2. You'll need to use your answers to Q3. (*Hint*: `dmy()`.)
   - Lastly, check if there is any cleaning required. If so, clean the data now.

   Output the first 10 rows, and the dimensions, of your clean, tidy and tame data set.

5. Making sure you set the seed correctly (according to the Deliverable Specification), use `sample_n()` to choose a random sample of 450 of the archers. Make sure they are ordered by the ID number and then output the first 10 lines of the dataset and the dimensions of the data set.

   Note

   Use this random subset from Q5 for the remainder of the assignment.

6. Now let's get on with some analysis. Add a new column to your tidy dataset recording the archers' accuracy:

   - `acc`: as the ratio of number of hits to number of shots

   Describe what type of variable this new column represents ("Quantitative Discrete", "Quantitative Continuous", "Categorical Nominal", "Categorical Ordinal"). Is the data type correct in the tibble? (Explain your answer.) If it is not correct, make sure you change it.

   Output the first 10 rows, and the dimensions, of the data set.

7. Use `inspect_num()` to display the summary statistics for the numerical values in your dataset. What are the means for the number of days of experience and the accuracy?

8. We'd like to compare the accuracy to the amount of experience of the archers, but they are measured in very different units. So use `predict()` with the commands from the `caret` package to standardise the numerical variables in your data set. *(Hint: page 3 in Module 5.)*. Output the first 10 rows, and the dimensions, of the standardised data set.

9. With the standardised experience and accuracy variables from Q8:

   (a) Which variable has the higher mean?

   (b) Which variable has the higher median?

   (c) Which of the variables has the higher inter-quartile range?

   (d) Plot histograms of each variable and calculate the skewness. (Make sure you use the correct skewness function from Prac 5.) Describe the shape of the histogram in the context of the data.

10. Returning to the non-standardised data from Q6, create a scatterplot of `acc` against `xp`. Put the independent/explanatory variable on the horizontal axis. Include a straight line of best fit on your plot. Does it look like there is a linear relationship between the two variables? (Provide some reasons for your answer.)

    Describe how this relationship coheres or conflicts with what you found in Q9.

11. We would like to fit a linear model to this data, and so we will first apply a Box-Cox transformation to the **response** variable.

(a) Use `BoxCoxTrans` to obtain an estimate of λ. (Extend the range of the search for $-10 \leq \lambda \leq 10$, in steps of 0.1.) What is the estimated λ?

(b) Apply the transformation to create a new column called `acc_bc` on the right of your dataset.

Output the first 10 rows, and the dimensions, of the data set.

12. Produce a scatterplot of the Box-Cox transformed data (with a line of best fit), as well as a histogram and the skewness. Write 2–3 sentences about this output.

13. (a) Write down the general equation of a linear model for the entire population (not just your sample). Make sure you define all of the notation you introduce. *(Hint: this equation should include the error terms.)*

(b) Also write down the formula for the line of best fit for your specific sample (with standard notation for the coefficients). Again, make sure you define any new notation.

(c) Now build a linear model in `R`, and use the output to find estimates for the model coefficients, and write down your sample line of best fit.

14. Using the material in Module 6 (p. 10–13), check if our model satisfies the 4 assumptions for a linear model. You will need to write some text explaining your conclusions here. Make sure you identify at least one possible problem with the **Independence** assumption.

- *(Note that we are going to use a linear model regardless of any problems that you find in the assumptions, but it is always good to highlight any shortcomings of the model so the client knows about them.)*

15. Use your model to predict the accuracy for an archer with experience of 2, 5, 10, 15, 20 and 25 years. Make sure you provide the correct interval as well. *(Hint: you will need to transform your predictions and intervals back into the scale of the original variables.)*

16. Write a paragraph or two describing what you have found about how long it takes to become a good archer, and any observations or conjectures that you have. (Everybody's data will be different so there is no right or wrong answer here, as long as you justify your claim with reasonable arguments.)

That's enough for this report. We might investigate this data further in Assignment 3. (Then we can charge the client more money for another deliverable!)

4 Submission

You must submit your assignment via MyUni. Do not email it to the teaching staff. Detailed instructions are on the assignment submission page in MyUni. Make sure that all your output is relevant to the questions being asked.

5 Deliverable Specifications (DS)

Before you submit your assignment, make sure you have met all the criteria in the **Deliverable Specifications (DS)**. The client will not be happy if you do not deliver your results in the format that they've asked for.