MATHS 7107 Data Taming
Assignment 3

Trimester 2 2024

Source: Jay's Classic Movie Blog

1 Background

With your help on their previous project, the archery club has been busily recruiting new members. They now have 90,000 members spread over 3 different venues. With such a huge membership, it's become clear that Robin Hood's Band of Merry Men are joining the club to keep their archery skills up to scratch, and this has attracted the attention of the Sheriff of Nottingham.

A new archer has applied to join the club. He's just a young lad, barely out of school, but he gave a demonstration of his archery skills that was very impressive: he hit the target with 112 arrows out of 116 total shots. Even though he has no police record, the Sheriff believes that this boy may be a Merry Man, as he saw the boy wearing a green outfit near his home in the forest.

The club has provided you with 3 data sets from a more recent tournament, featuring all their members. Using this data, try to help the Sheriff determine if this new archer is indeed a Merry Man. Conveniently for you, the Sheriff has just started using `R` and `R Markdown`, so he wants your report as a PDF generated using `R Markdown`. He studied Data Taming last trimester, so he wants you to only use commands from the course, so that he 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 he can see what `R` code you used to generate your output. But of course, he 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 `merry_0.csv`, `merry_1.csv` and `merry_2.csv` (one from each of the venues). Each dataset contains 3 columns:

- `RHBMM`: if this person is known to be a Merry Man then this is `1`, otherwise it is `0`.
- `Accuracy`: the accuracy the archer achieved in the tournament.
- `AGE`: the age bracket the archer falls into `youth`, `middle` or `senior`.
- `DRESS`: the colour of clothing the archer wears, `black`, `red` or `green`.
- `Home`: where the archer lives, in a `city` or a `forest`.
- `Jail`: `yes` or `no` indicating whether or not the archer has spent time in jail.

Each dataset has data on 30,000 archers. Luckily, the data has already been cleaned and so there should not be any missing or erroneous data. (If you find any problems with erroneous data, then report this to the archery club immediately, so that their data cleaner can be fired.)

IMPORTANT!

If you remove any data, then 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

To help the Sheriff, we will analyse the data for the known Merry Men, and make a prediction about whether this new archer is a Merry Man or not.

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. 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.
3. Now it's time to tame our data. But since we are going to fit a logistic regression model, we need to modify our requirements a little bit.

- Make sure that all column names are in snake case.
- Convert the Merry Men status to a `<fct>` data type, with `yes` for `1` and `no` for `0`.
- If you have identified any Categorical Ordinal variables, store them as a `<fct>`.
- Make the remaining variables conform to the Tame Data conventions in Module 2 (page 3).

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

4. Setting the correct seed, split your data into a training set (with 20,000 rows) and a testing set, with the remaining rows. Output the first 10 lines of each dataset and the dimensions of each data set.

5. Fit a logistic regression model to your training data, with the Merry Men status as the response and all other variables as the predictors. (Just use them individually, don't include any interaction terms.) Output the summary of the model.

6. Use the command `model_matrix()` on the `DRESS` and `AGE` variables of your training data to see what happens to them when we fit a model. (See pages 2 – 6 in Module 7.)
   (a) How many new variables have been introduced? (Make sure you explain your answer here.)
   (b) What are the reference levels for `DRESS` and `AGE`?

7. Since we are using general linear models, the model summary in Question 5 describes linear geometric objects, where the dimension of the geometric object is determined by the number of continuous predictors. We have only a single continuous predictor so our model describes a set of lines. How many lines are described by the model in Question 5? Make sure you give some justification for your answer.
   - *(Hint: see the Week 8 seminar and pages 12 – 16 of Module 7. The model summary and the* `model_matrix()` *outputs should help.)*

8. Now it is time to get serious with our data. There may be some interactions between the variables in the data set, so fit a new model to your training set using all the individual variables and all the second-order interaction terms. Use `Anova()` to find the p-values for each of the variables. Identify all interaction terms that meet the 99% significance level.
   - *(Hint: if you have three predictors x_1, x_2, x_3, then the second order interaction terms are $x_1x_2,\ x_1x_3,\ x_2x_3$. There is an easy way and a hard way to do this — see the Reminder sheet for the easy way.)*

9. We'll now apply **backwards stepwise regression**. As we learned in Module 7, best practice is to only remove terms one-by-one starting with the least significant. However, the Sheriff has told us that he doesn't believe there should be any interactions between the archers' criminal record and their other characteristics.
   (a) So first fit a new model with ALL the individual variables and all interactions except for those interactions involving the `Jail` data. Show the `Anova()` output.
   (b) Then ONLY looking at the interaction terms, continue with step-by-step **backwards stepwise regression** to find a model where all interaction terms meet the 95% significance level. At each step, identify the interaction term that you will remove, and why you will choose that one. Then show the resulting `Anova()` after you fit each model.
      - The **"principle of marginality"** tells us that a variable shouldn't appear in an interaction term if we don't have the variable appear by itself.

   (c) Finally, now focus on the individual terms and finish the **backwards stepwise regression** so that all terms (individual terms and interaction terms) meet the 95% significance level. At each step, identify the variable that you will remove, and why you will choose that one. Then show the resulting `Anova()` after you fit each model.

10. (a) Which interaction terms are significant in your final model?
    (b) Thinking about the context of the data, provide some reasonable hypotheses for why those interaction terms might represent real effects (and are not just statistical noise).

11. So we have now fit a logistic regression model for the **log-odds**, which has the general form:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{r}_i$$

where $\hat{r}_i$ is an estimated function of the predictors. Write down the general form of $\hat{r}_i$ for your final model in Question 9. Keep the coefficients as pronumerals for now, so it should look like:

$$\hat{r}_i = \hat{\beta}_0 + ...$$

Be sure to define all variables in your equation.

12. Looking at Question 11, the geometric situation is slightly more complicated now than in Question 7, although our model should still produce a set of lines.

    (a) How many lines does your final model describe? Make sure you provide some justification for your answer.

    (b) Are the lines all parallel? If not, explain why not.

13. Now output the summary of your final model showing the estimated coefficients, and use that to write $\hat{r}_i$ with all the estimated coefficients replacing the $\hat{\beta}_j$ pronumerals.

14. What is our estimate for the **log-odds** of an archer being a Merry Man:

    (a) who is middle aged, lives in the forest, wears red clothing and has been to jail?

    (b) who is quite old, lives in the city, usually wears black clothing and has never been to jail?

15. Now apply your final model to the testing data. Produce a new tibble containing the true classes, the predicted classes and the prediction probabilities. Output the first 10 lines of this tibble and the dimensions of the data set.

16. Now we need to evaluate our model.

    (a) Find the confusion matrix and the accuracy of the model.

    (b) If "being a Merry Man" is classified as a success, find the sensitivity and specificity of our model. (*Hint: make sure you calculate the values yourself, as R may not choose the right level.)*

    (c) Plot the ROC curve. You might want to add the following code to your `autoplot()`

    ```
    + geom_vline(xintercept=...)  + geom_hline(yintercept=...)
    ```

    to help you identify which is the correct ROC curve.

    (d) What is the AUC of this ROC curve?

17. Finally, let's see what we think about this new archer. Based on your model, do you predict that the new archer is a member of Robin Hood's Merry Men? Write some text to interpret your results for the Sheriff, and make sure you give the probabilities of your predicted class.

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.