MATHS 7107
Data Taming
Practical 9

Revision

1 Preliminaries

- Set up a project in RStudio
- Download the `test_runs_original.csv` file to a `data` subdirectory of your project directory
- Now load the packages
  - `tidyverse`
  - `tidymodels`
  - `car`
- Read in the data. You can call it whatever you like, but in the commands below we have assumed the data will be called `test_runs`.

1.1 Aim of today's prac

We are going to build a model to predict the batting average of a professional cricket player. But first we'll need to clean up the data, and then we'll have a look at some of the statistics in the dataset. We're going to use the dataset `test_runs_original.csv`, which has the following columns:

1. No column name (although `R` will give it the name `...1`) — the player's row number in the table.
2. `Player` — the player's name, along with the country that they played for.
3. `Span` — the years over which the player was playing professional cricket.
4. `Mat` — the number of matches they played in.
5. `Inns` — the number of innings that they batted in. (A cricket test match typically contains 2 innings.)
6. `NO` — the number of innings where they were "not out" (ie. they played all the way to the end of the innings).
7. `Runs` — the total number of runs they made in all of the innings in their career
8. `HS` — the highest score they recorded in a single innings. Some of the entries have a "*", and we are unsure what that means.
9. `Ave` — the average number of runs per innings over their career
10. `'100'` — the number of innings where they scored at least 100 runs. This is called a "century".
11. `'50'` — the number of innings where they scored at least 50 runs.
12. `'0'` — the number of innings where they did not score any runs.

2 Taming the data

Questions:

1. Extract the country code from the player variable and create a new variable called Country. Hints:
   - You can use the regular expression:

     `\\((.+)\\)`

     where the double backslashes "escape" the round brackets, because they are special characters in regular expressions. The "dot" stands for "any character", the "+" says get as many as possible.
   - To get rid of the `ICC/` you can use the regular expression:

     `(.+)/`
2. Remove the country code from the players name.
3. Extract only the number from the `HS` (Highest score) variable.
4. Add a new column called `Years` which gives the number of years in `Span`. Put this new column just to the right of `Span`.
5. Change the following variable names
   - `...1` to `rownum`
   - `100` to `Centuries`
   - `50` as `Fifties`
   - `0` as `Zeros`
6. Recode the relevant variables as factors, ordered factors and integers

3 Descriptive statistics

Questions:

7. Who scored the highest test score? and from which country?
8. Compare the batting averages from each country
9. Create a bar graph for the proportion of centuries from each country.

4 Model

Questions:

10. Build up a model to predict the average score by using the following predictors.
    - Mat
    - NO
    - HS
    - Centuries
    - Fifties

- Zeros
- Country
- Years

Make sure you optimise your model by performing "backwards stepwise regression".

11. Check the assumptions for your optimised model.
12. Predict the batting average of a player with the following statistics

    - 85 matches (Mat)
    - 17 Notouts (NO)
    - 27 Centuries
    - 36 Fifties
    - 12 Years

    Also include prediction and confidence intervals at the 99% level.