STA4003 Project

Due date: December 15, 2023

- Outstanding projects will be invited to give a presentation on Dec 7, 2023. Students who have given a presentation can receive maximum 10 bonus points in their final exam.
- Students who want to present their work need to submit the project by Dec 1, 2023. Submissions after Dec 1 will not be invited for presentation. All students can revise their work before Dec 15, 2023.
- The submitted codes must be clearly written in a `R` file.
- A report to describe your analysis is required.

1 Background

In this project, we will analysis a dataset about horse racing. Let's have a brief introduction of horse racing. In a particular game, there are 14 horses racing. Before a particular time t_{final}, people are allowed to bet which horse can win the game. Let $b_i(t)$ be the total amount betting on horse i at time t. Note that $b_i(t)$ is increasing before t_{final}. After the game, we have $b_i(t_{final})$ being bet on horse i for $i = 1, \ldots, 14$. If horse I wins the game, people who bet on horse I can get the dividend

$$d_I^f = d_I(t_{final}) = \frac{(1-\nabla)\sum_{j=1}^{n} b_j(t_{final})}{b_I(t_{final})}$$

for each \$1 bet, here $\nabla = 0.175$ is the percentage track-take. Note that the dividends

$$d_i(t) = \frac{(1-\nabla)\sum_{j=1}^{n} b_j(t)}{b_i(t)}$$

for horse i, $i = 1, \ldots, 14$, are known by all gamers at time $t < t_{final}$. As $b_i(t)$ is time varying, so does $d_i(t)$.

Now suppose we have some insider information and we believe that we know the "true" winning probability π_i of each horse i. Since we will only make a bet on horse i if the expected

return is greater than $1/\pi_i$, so one betting strategy is betting on horse i if $d_i^f > 1/\pi_i$. However, we don't know d_i^f at time we bet (t_{bet}). Let $b_i = b_i(t_{bet})$, $d_i = d_i(t_{bet})$, $f_i W$ be the amount we bet on horse i at t_{bet} and C_i be the amount bet on horse i by other parties after t_{bet}. Then we have

$$d_i^f = \frac{(1-\nabla)\sum_{j=1}^{n}(b_j + C_j + f_j W)}{b_i + C_i + f_i W}.$$

The unknown quantities here are C_i for $i = 1, \ldots, 14$. Obviously, the amount of C_i's affects the accuracy of the strategies that are based on the values at time t_{bet}. In this project, your task is to analyse the time series $C_{sum} = \sum_{i=1}^{14} C_i$.

2 Data

The datasets "data20XX.RData" with XX=14,15,16,17,18 are given. They all have the same set of column names, which are

- **ID**: It is of the form "yyyymmddrr", which means Year yyyy Month mm Date dd Race rr. Note that there are more than one race on each day and the number of races can be different on each day.
- **WIN_POOL.x**: The total amount in the pool at time t_{bet}.
- **WIN_POOL.y**: The total amount in the pool at time t_{final}. Hence C_{sum} is the difference between **WIN_POOL.y** and **WIN_POOL.x**.
- **WIN_TAKE.x**: $\nabla = 0.175$. It is the same as **WIN_TAKE.y**.
- **WIN_ODDS_i.x**: $d_i = d_i(t_{bet})$. If it is 0, it means that horse i actually was not in the race.
- **WIN_ODDS_i.y**: $d_i^f = d_i(t_{final})$. If it is 0, it means that horse i actually was not in the race.
- **WIN_MODEL_i.x**: "True" winning probability π_i. If it is 0, it means that horse i actually was not in the race. It is the same as **WIN_MODEL_i.y**.
- **WIN_TIME.y** The "yyyymmdd" part of **ID**.
- **WIN_NUMBER.y** The "rr" part of **ID**.

3 Tasks

In this project, you are required to forecast C_{sum} for each race in `data2018.RData`. Note that you MUST only use the information BEFORE t_{bet} to forecast the C_{sum} in a particular race. Let N be the total number of races in 2018, x_r be the true C_{sum} on Race r, $\hat{x}_r$ be your

forecast, and $f_{p,r}$ be your quantile forecast with probability $p = 0.95$. You should include the followings in your project.

1. (10 points) Describe clearly the model you used for forecasting x_r based on the information prior to the time t_{bet} for Race r. That is,

$$x_r = H(\mathcal{F}_{r,t_{bet}-}) + e_r, \tag{1}$$

   where $\mathcal{F}_{r,t_{bet}-}$ is the information prior to the time t_{bet} for Race r, H is some specific function you need to describe, and e_r is the error term.

2. (20 points) Compute the mean absolute percentage error MAPE described in Section 5.8 in the textbook "Forecasting: Principles and Practice, 3rd Ed" for you forecasts. Your codes must output the mean absolute percentage error in a variable `MAPE`.

3. (20 points) Compute the quantile score $Q_{0.95,r}$ for each Race r in 2018. And then report the average quantile score in a variable `QS`.

Please note the followings.

1. Your work will be evaluated by other dataset, namely "data2019.RData", that have the same set of columns of the given data set.

2. Only the given data set and the information provided in the project can be used. Don't use any other additional information in your analysis.