STA4003 Project Report

Yuzhou Peng

121090446

The project is divided into 3 major components: data preparation, model fitting and result output.

1. Data preparation

    (1) Combine the data from 2014 to 2017 in **Train_all**

    (2) Extract data of 1st, 2nd, 3rd ....11th race of the racing days and form data **train_all_i** and **test_i** where i = 1,2,3....11

2. Model fittting

    The model includes 2 models:

    (1) a linear model with **Csum** , **WIN_POOL.x** being the response variable (y) and explanatory variable (x).

    (2) a vector autoregression (**VAR**) model

    For data of i-th race, we fit the model as follow.

    Step 1: Fit a **TSLM** model to the first 37.5% days (rounded by **floor()**) of the i-th races and predict the first 30 days data in the test set. If the test set have fewer days then 30, predict all data using this model.

    Step 2: Fit a vector autoregression model to the first j (j > 30) days in test set and predict data on day j+1. And thus iteratively generate all prediction of data after 30th day.

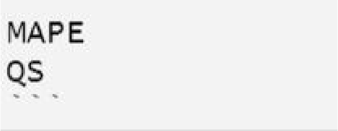
    (3) The result is stored in **result_i** for i-th race. **result_all** is the combined set of all **result_i**, containing **true_value**, **forecast_value**, **upper_quantile** (0.95 quantile) and **APE** (absolute percentatage error).

3. Result output

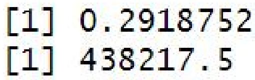
    (1) MAPE is calculated by the mean value of **result_all$APE**.

    (2) 0.95-quantile score is stored in **QS**.

    Note: The calculation of QS is based on the function **quantile_score** (consisting with our textbook) with input containing 0.95 quantile (**upper_quantile**) and true value of Csum (**true_value**)



~~~
MAPE
QS
```
~~~



```
[1] 0.2918752
[1] 438217.5
```