# STA4003_Report

---

## Note: You can find Direct Answers of Question 1, Question 2, Question 3 in Part III

## 1. Calculate $C_{\text{sum}}$

We denote

$$P(t) = \text{Total Amount in the Pool at } t$$

For each day each race, Note that

$$d_i(t_{\text{bet}}) = \frac{(1-\nabla)P(t_{\text{bet}})}{b_i} \qquad d_i(t_{\text{final}}) = \frac{(1-\nabla)P(t_{\text{final}})}{b_i + C_i + f_i W}$$

We assume each time we do not make any bet (or our bet is negligible compared with the pool), thus $f_i W = 0$ we can compute $C_i$ by:

1. Compute $b_i$:

$$b_i = \frac{(1-\nabla)P(t_{\text{bet}})}{d_i(t_{\text{bet}})}$$

2. Use $b_i$ to compute $C_i$:

$$C_i = \frac{(1-\nabla)P(t_{\text{final}})}{d_i(t_{\text{final}})} - b_i = (1-\nabla)\left(\frac{P(t_{\text{final}})}{d_i(t_{\text{final}})} - \frac{P(t_{\text{bet}})}{d_i(t_{\text{bet}})}\right)$$

Then, we can have $C_{\text{sum}}$:

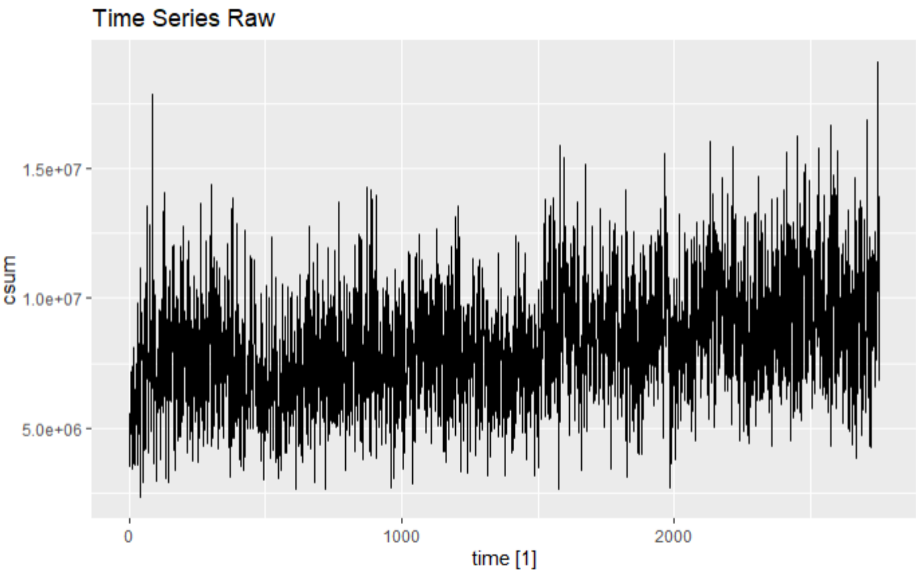
$$C_{\text{sum}} = \sum_{i=1}^{14} C_i$$

# 2. Data Analysis

We use data2014 to data2017 (four data sets) as our training sample and data2018 as the testing sample.

## 2.1 Differencing to Stationarity

We plot the time series:



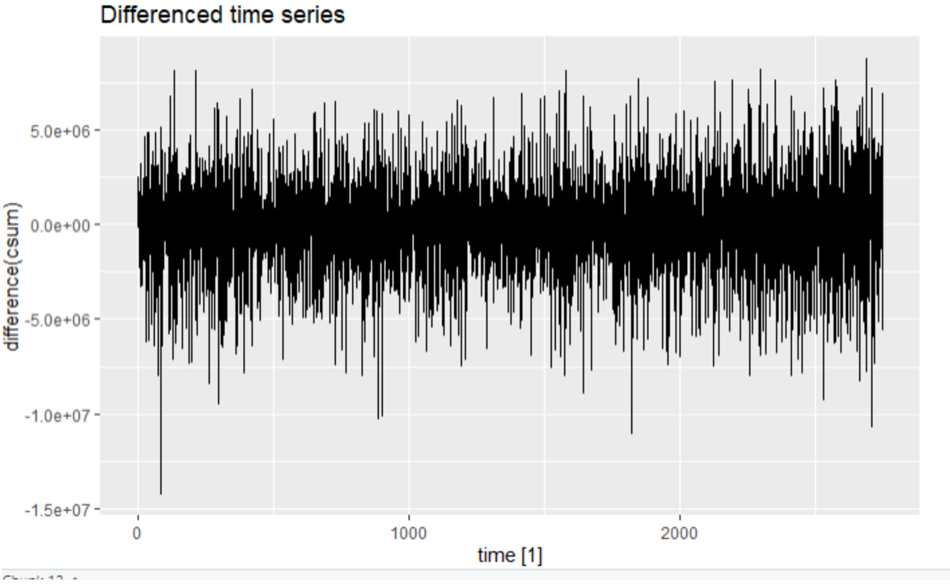

Intuitively, it is non-stationary. To justify this, we perform different unitroot testings, the result is showing as:

1. KPSS Test: The KPSS statistic is 12.84975 and the p-value less than 0.01, strongly rejected the null hypothesis that this series is stationary.
2. ndiff Test: The returned value is 1, indicating one difference is required.

Inspired by stationarity, we take differences in the time series. We try $d = 1$, We have the testing results:

1. KPSS Test: The KPSS statistic is 0.004319879 and the p-value larger than 0.1, thus we cannot reject the hypothesis that this is a stationary time series
2. ndiff Test: The returned value is 0, indicating no seasonal difference is required.

We plot the differenced series:

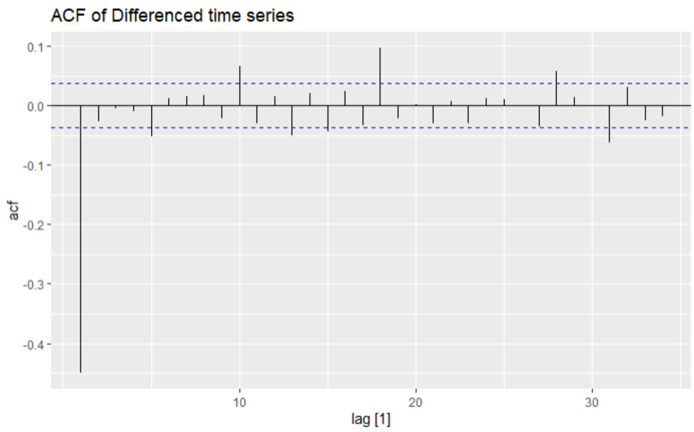


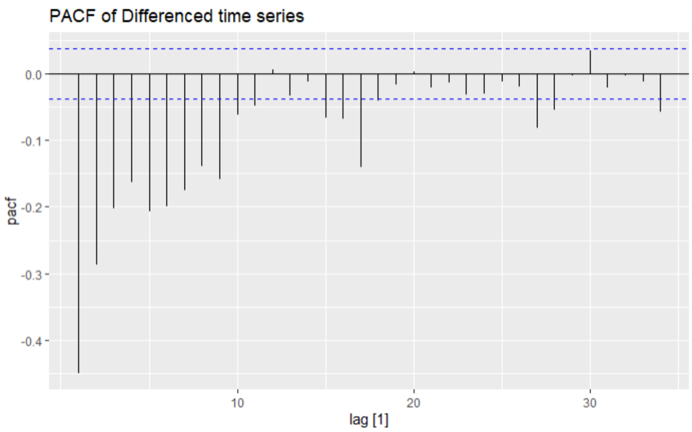

Intuitively, this "looks like" stationary and it passes the two tests. We may assume it is indeed stationary.

Therefore, since our main model is ARIMA(p,d,q), we have established the value $d = 1$.

## 2.2 Determine $p, q$

Recall the Theorem from lecture: If $x_t$ follows MA(q), we have $\rho(h) = 0, \forall h > q$. If $x_t$ follows AR(p), we have $\phi_{hh} = 0, \forall h > p$. For ARMA(p,q) with $p > 0, q > 0$, neither $\rho(h)$ and $\phi_{hh}$ will cut off at finite lags. This enlightens us to check the ACF and PACF plots:




From ACF, we get insights for $q$: Note that ACF is generally cut off (statistically indistinguishable from 0) for lag larger than 5, we primarily set $q \in [0, 6]$. From PACF, for lag larger than $10$, we have PACF insignificanly distinguishable from $0$ generally, we primarily determine $p \in [0, 10]$.

Now, we consider the range inspired by the plots for p,q and we formally conduct AIC to further determine the optimal p,q. The pair p,q that minimizes the AIC is given by $p = 5, q = 7$. We can also use BIC and getting $p = 5, q = 7$. Note that BIC is giving the same pair as AIC Therefore, we choose

$$p = 5 \qquad q = 7$$

## 2.3 Use of $\pi_i$

It is expected that, if $d_i > \frac{1}{\pi_i}$, there will be more investors who want to make the bet, thus expectedly the $C_{sum}$ will increase correspondingly. We count the number of $\pi_i d_i > 1$ where $i = 1, \cdots, 14$ as the indicator of $C_{sum}$, denoted as x_count. Therefore, we first regress our $C_{sum}$ on $\sum_{i=1}^{14} \mathbb{1}(\pi_i d_i - 1)$. Then, we use the residual to do time series analysis

# 3. Modeling and Forecast

## 3.1 Training (Question 1)

By Previous Analysis, we use the ARIMA(p,d,q) model. For race 1, we have

$$p = 5 \qquad d = 1 \qquad q = 7$$

Therefore, we have

$$y_t = \beta_0 + \beta x_t + \eta_t + e_t$$

where the $x_t$ is the external regressor

$$x_t = \sum_{i=1}^{14} \mathbb{1}(\pi_i^t d_i^t - 1)$$

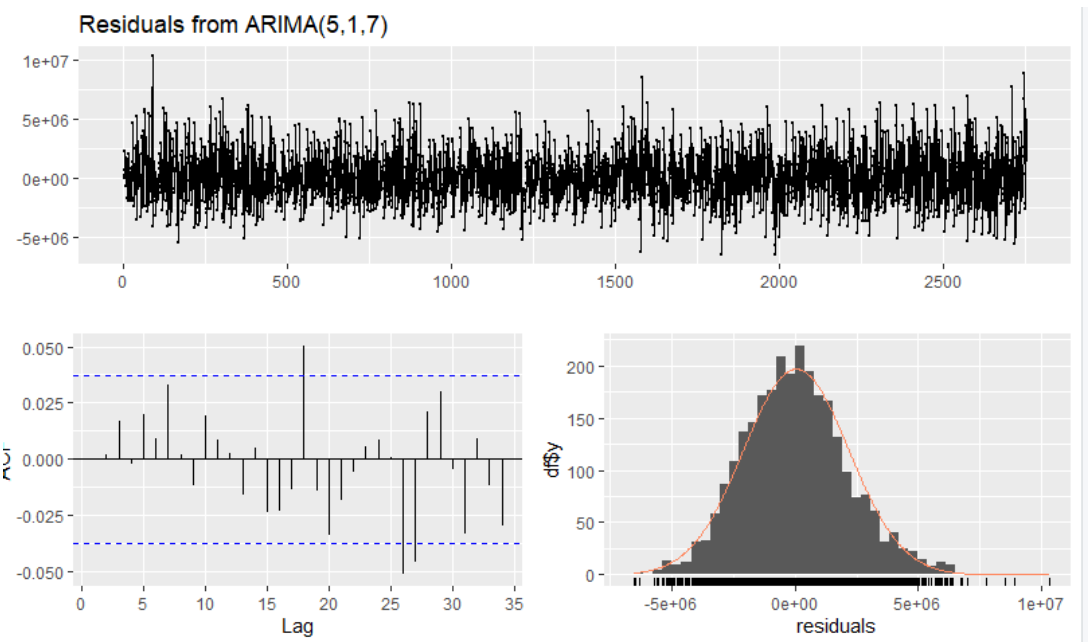
and

$$\eta_t = \phi_1 \eta_{t-1} + \cdots + \phi_5 \eta_{t-5} + w_t + \theta_1 w_{t-1} + \cdots + \theta_7 w_{t-7}$$

and $e_t$ is the error term.

We train this model using data from 2014 to 2017, and we train this using arima.

The residue analysis is:

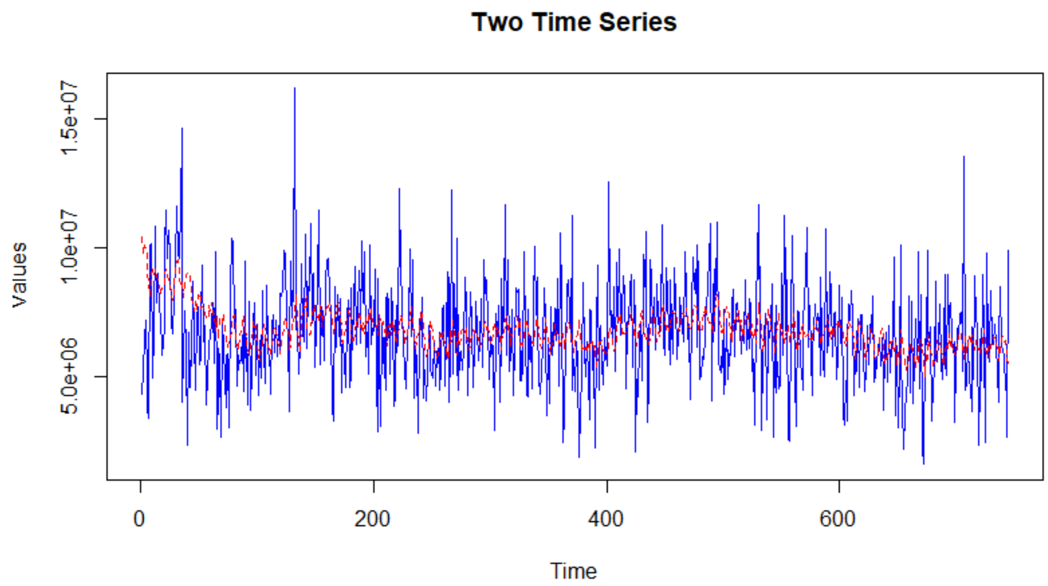


## 3.2 Forecasting (Question 2)

We use the data from 2018 to do the forecasting. We have the following MAPE:

$$MAPE = 22.5628\%$$

The result is given by




## 3.3 Forecasting Quantile (Question 3)

To obtain the forecast 95% quantile, we obtain the 90% confidence interval's upper bound of the forecasted value for every point in the test time series, stored at `p95_value` variable.

The Quantile Score is given by

$$Q(f, y) = \begin{cases} 2(1-p)(f-y) & y < f \\ 2p(y-f) & y \geq f \end{cases}$$

where $p = 0.95$.

The average 95% forecast quantile is given by

$$Q = 456246.2$$

This is stored at QS.