NFL Time Series Prediction

Angelica

5/16/2022

Introduction

The National football league (NFL) is a 6.5-billion-dollar brand, gaining international popularity by the year. In the last year alone, (2020 – 2021), the NFL grew in popularity by 14%. Additionally, the NFL adds 270 million to the gambling industry. Needless to say, understanding the analytics of football could score you a heavy paycheck.

The NFL consists of two divisions, the American Football Conference (AFC) and the National Football Conference (NFC). Currently, each division has 16 teams, for a total of 32 teams. For this project, I recreated the NFL data from the past 30 years (1991-2021), and the idea is to build a prediction/forecasting model of the number of wins for a particular team of our choice within the two NFL divisions and then compare the team's forecast predictions. For more information about NFL teams, see this link here; https://www.nfl.com/teams/.

Data source: https://www.nfl.com/standings/

Notes:

This is only regular season wins and losses, playoff games are not included. Thus, the forecasting will be on number of wins in a regular season.

Dataset Description

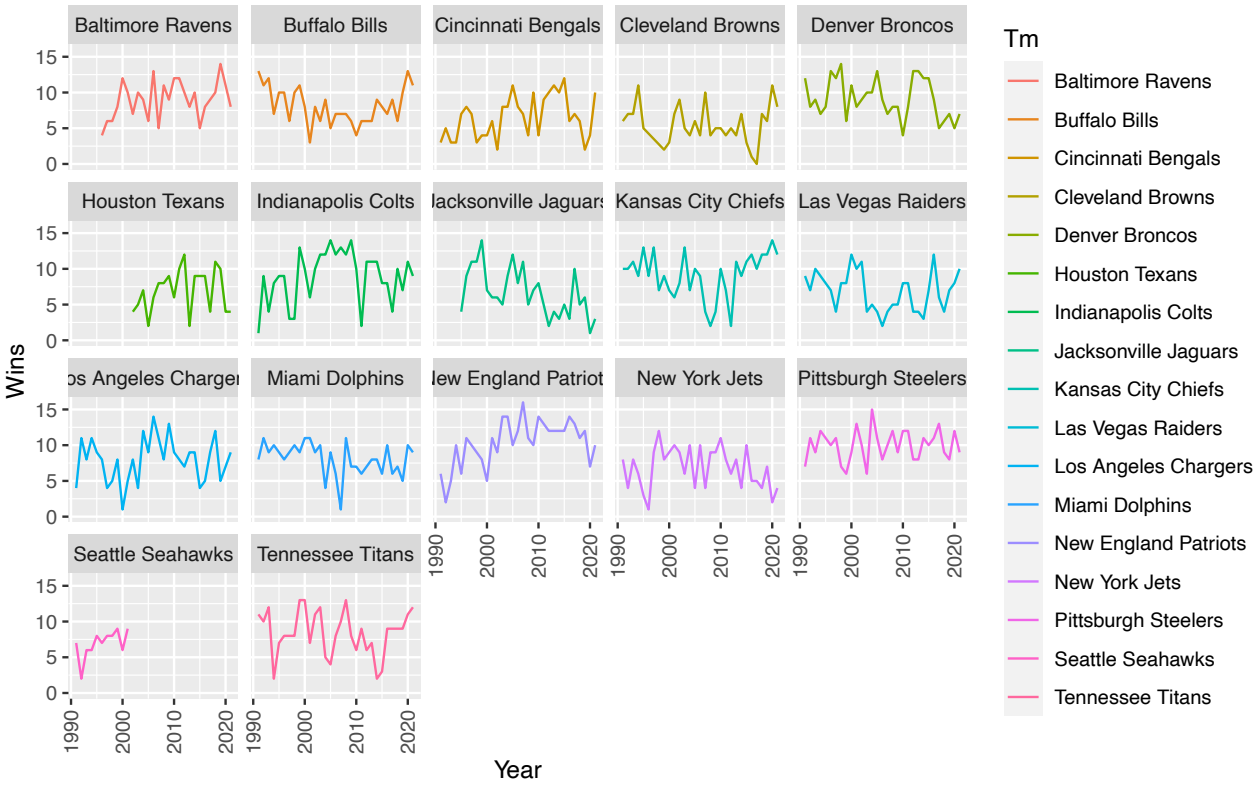
There are 13 variables:

Variable	Description	Data Type
Tm	American Football Team	chr
W	Games Won	num
L	Games Lost	num
W-L%	Win-Loss Percentage of Team	num
PF	Points Scored by Team	num
PA	Points Scored by Opposition	num
PD	Points Differential(PF - PA)	num
MoV	Margin of Victory	num
SoS	Average quality of opponent	num
SRS	Simple Rating System(MoV+SoS)	num
OSRS	Offensive SRS	num
DSRS	Defensive SRS	num
Year	Year of the NFL season	num

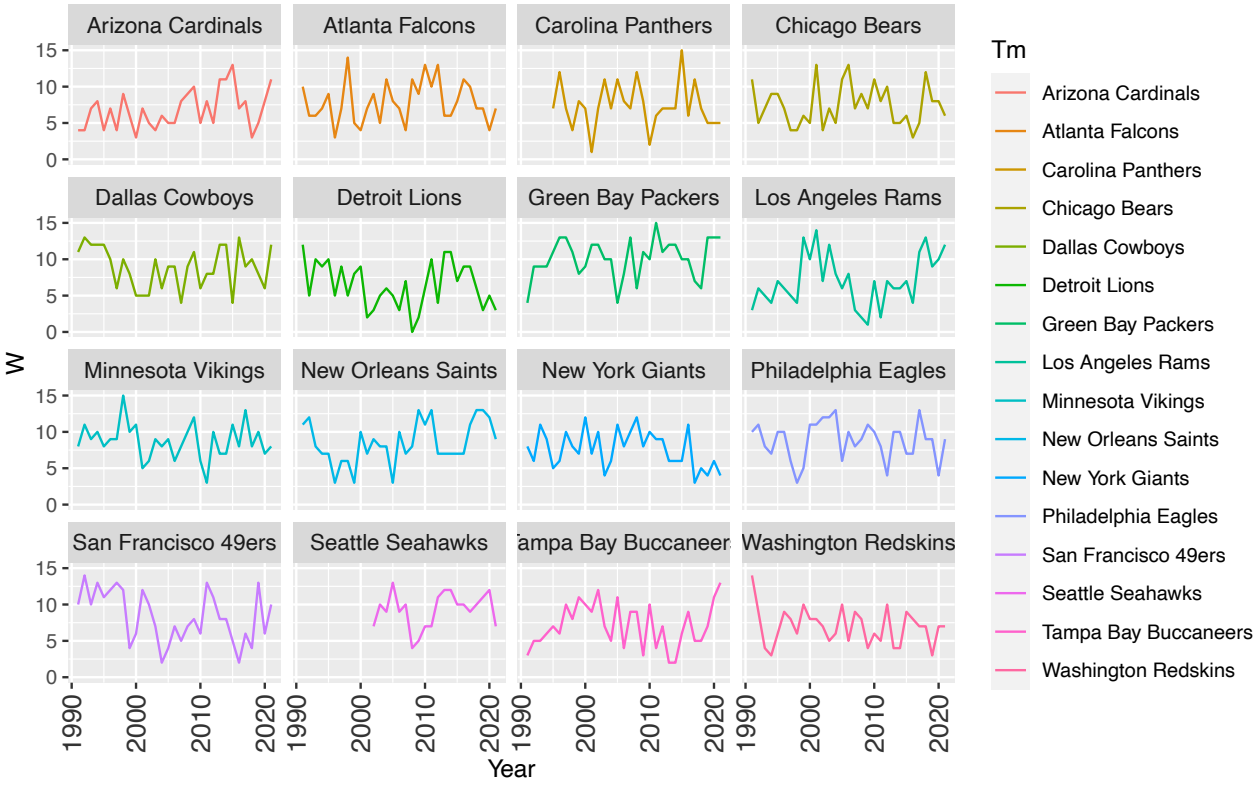
Exploratory Analysis

Starting with an exploratory analysis of the overall wins of each team in their respective NFL division. The time series plots from this analysis let us proceed with the team selection.

Overall Wins on all teams in AFC football division



Overall wins on all teams in NFC football division



Predictive Strategy

Prediction Problem: Create a model to predict the number of wins for a particular NFL team in the next 2 years (2022-2023).

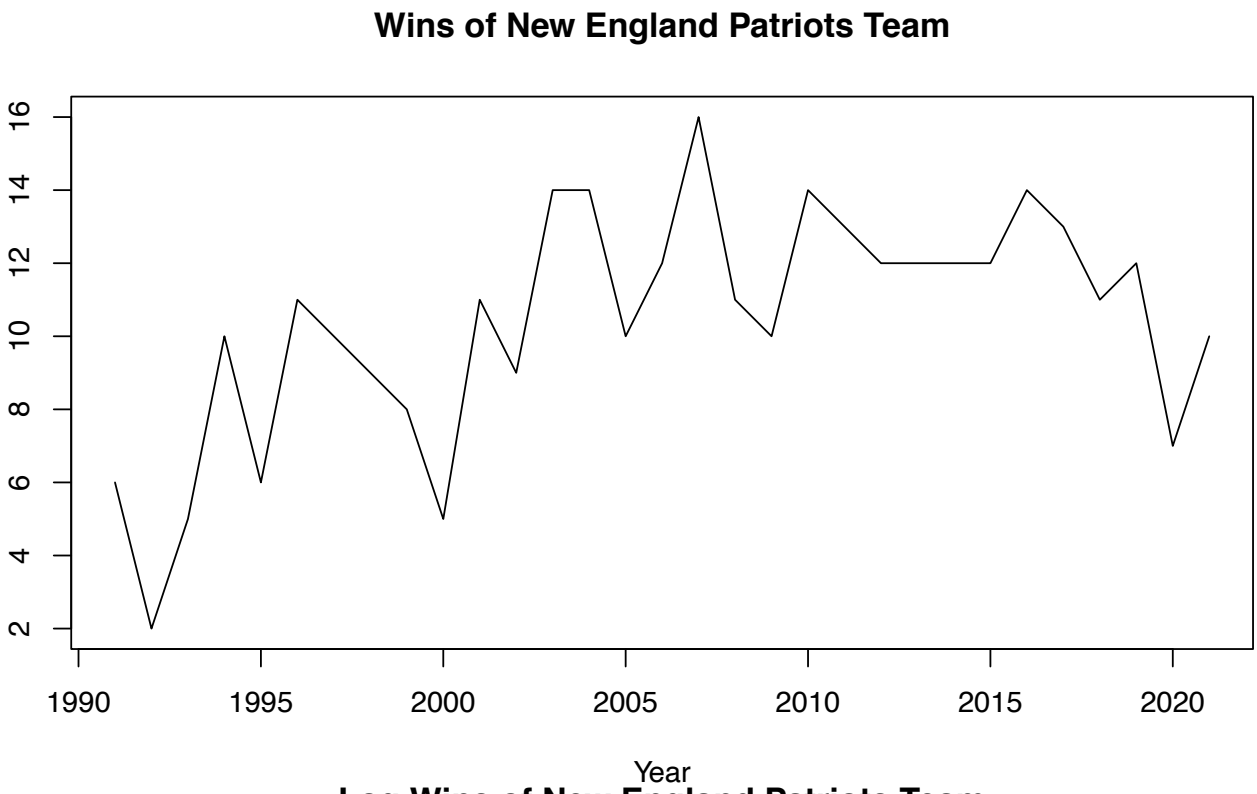
Description of the Strategy

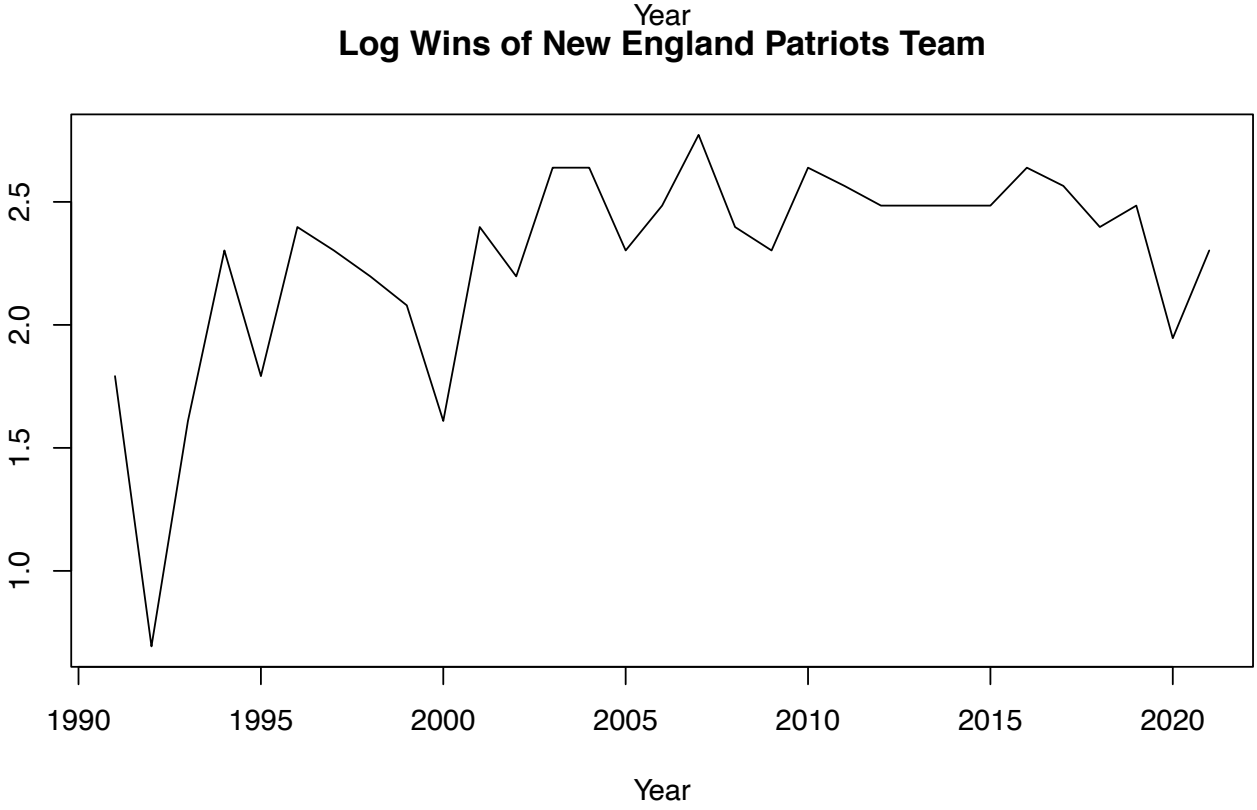
1. Plot the dataset and remove trend or seasonality if it is needed.
2. fit the best model implementing different forecasting methods.
3. Residual diagnosis to check best model assumptions.
4. Forecasting the number of wins for the next 2 years based on one NFL team.

Analyze Time Series on One Specific Team

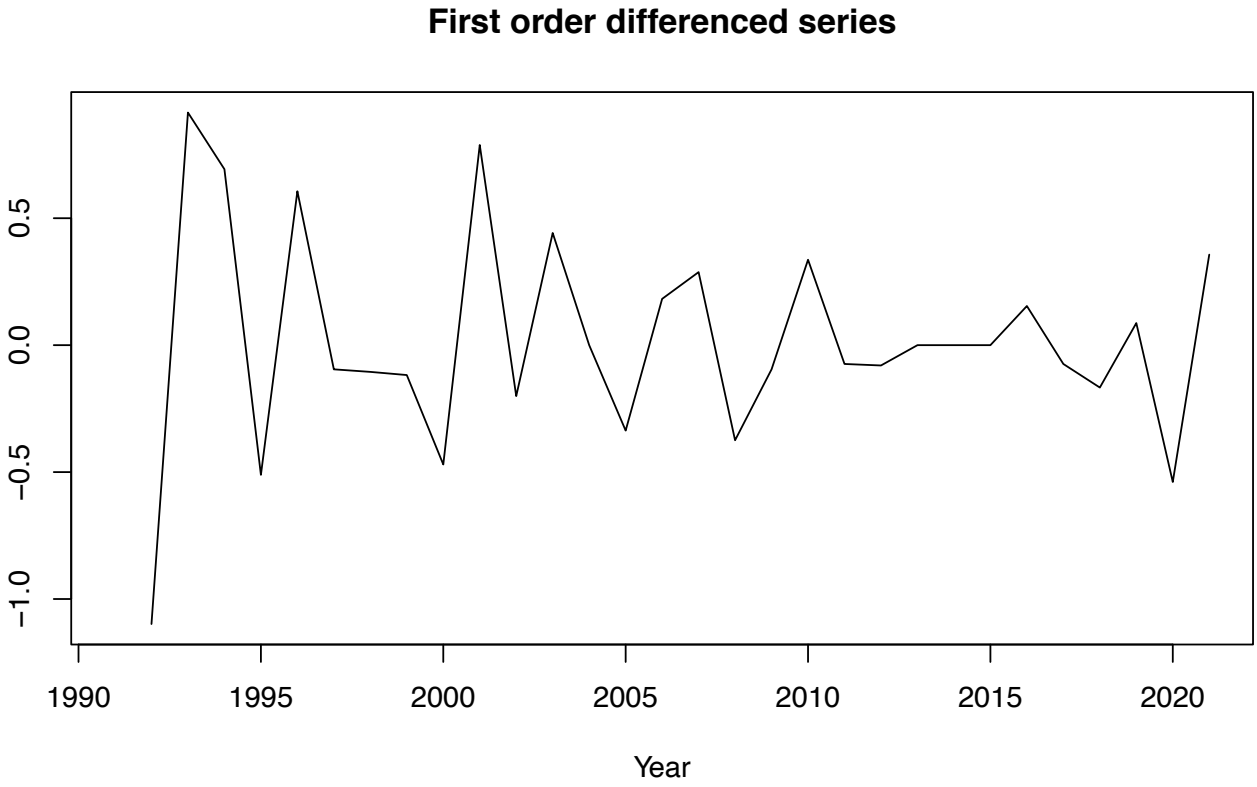
Based on the overall wins on the NFL teams, the New England Patriots is the team for time series analyses. The main goal is to build the best model and create a forecasting comparison with the final result.

Transformation of the New England Patriots' Win data

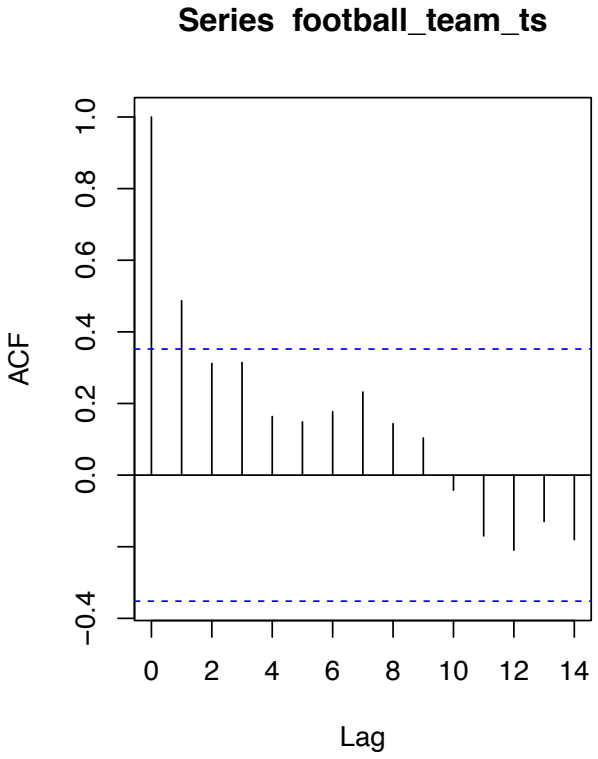


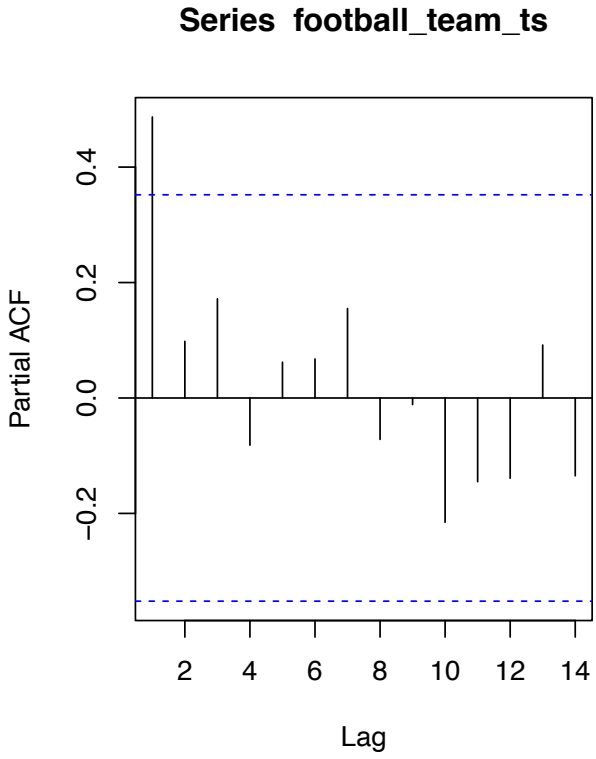


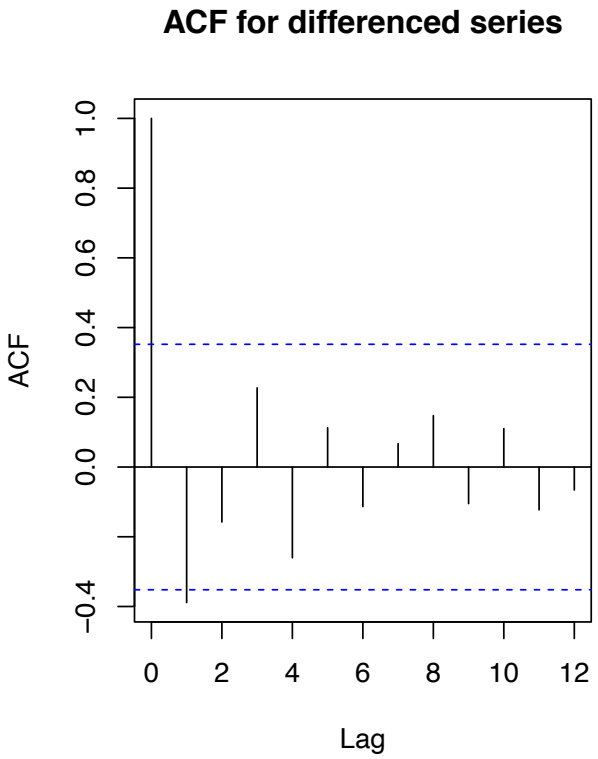
We can see a moderate increasing trend from 1992 to 2006. Log is an appropriate transformation to stabilize the data from the increasing trend over that period.

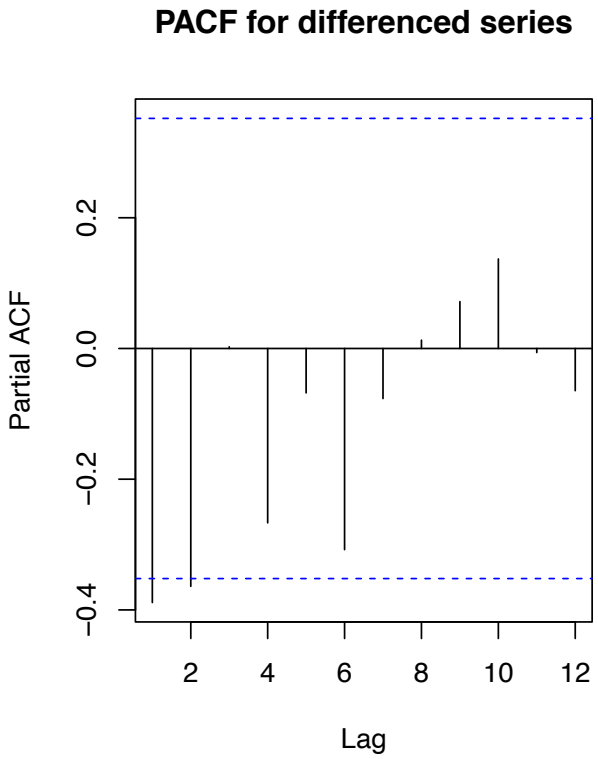


It appears that trend have been removed after applying First order differenced series. The differential order makes the time series stationary.









Based on the AIC resulting from the auto.arima function, ARIMA(0,1,1) could be the best fit model for this particular dataset. Next step, try a few models to confirm this assumption.

Fitting the best ARIMA model

AR(1)

```
##
## Call:
## arima(x = football_log, order = c(p = 1, d = 0, q = 0))
##
## Coefficients:
##          ar1  intercept
##       0.4876     2.2577
## s.e.  0.1557     0.1225
##
## sigma^2 estimated as 0.1292:  log likelihood = -12.41,  aic = 30.81
```

AR(2) model

```
##
## Call:
## arima(x = football_log, order = c(p = 2, d = 0, q = 0))
##
## Coefficients:
##          ar1     ar2  intercept
##       0.4108  0.2018     2.2224
## s.e.  0.1797  0.2474     0.1646
```

```
##
## sigma^2 estimated as 0.1261:  log likelihood = -12.09,  aic = 32.17
```

ARIMA(0,1,1)

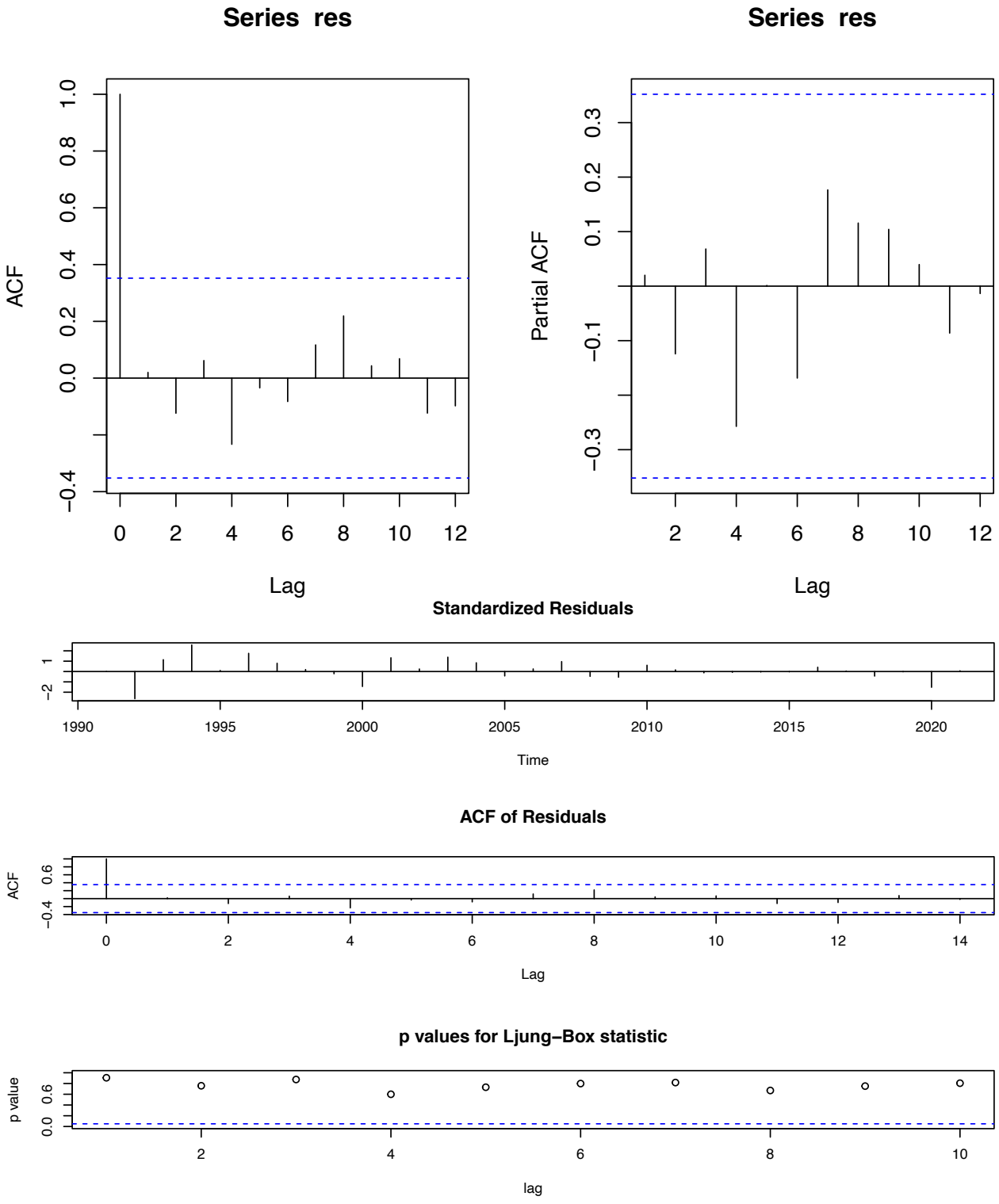
```
##
## Call:
## arima(x = football_log, order = c(p = 0, d = 1, q = 1))
##
## Coefficients:
##           ma1
##       -0.6087
## s.e.   0.1493
##
## sigma^2 estimated as 0.1273:  log likelihood = -11.88,  aic = 27.76
```

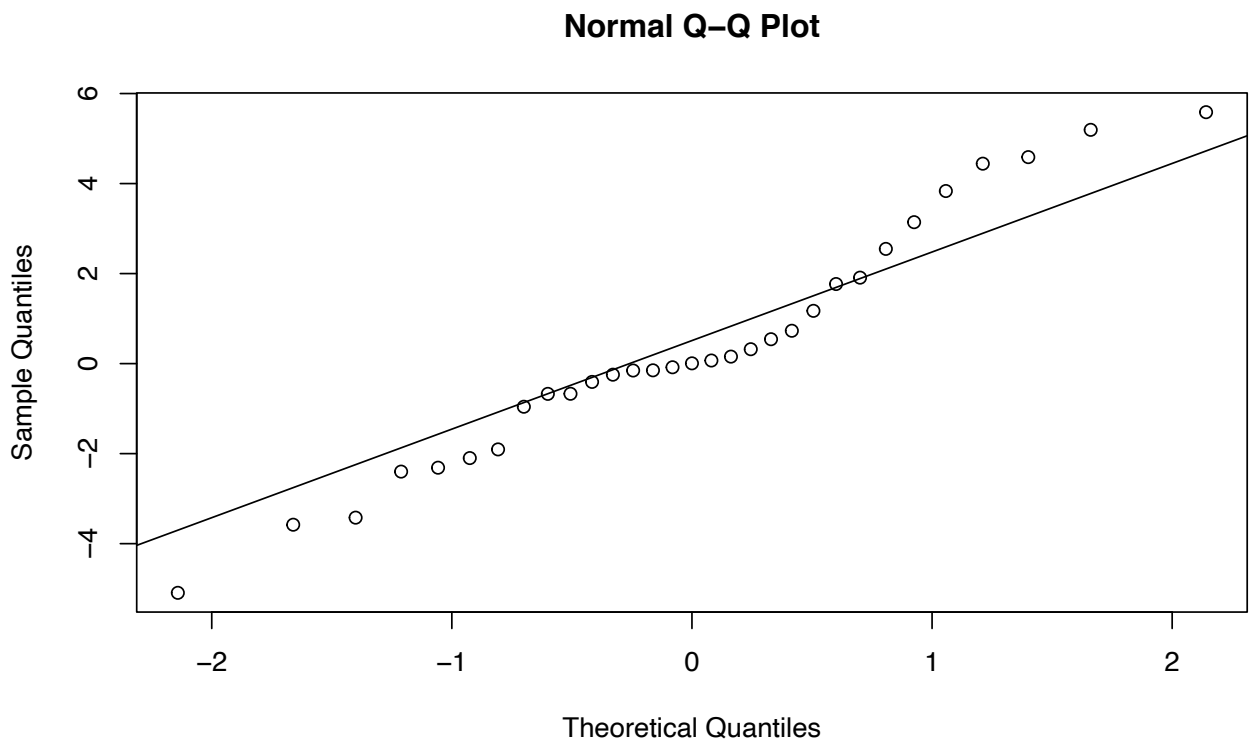
ARIMA(1,1,1)

```
##
## Call:
## arima(x = football_log, order = c(p = 1, d = 1, q = 1))
##
## Coefficients:
##          ar1      ma1
##       0.0249  -0.6257
## s.e.  0.3422   0.2708
##
## sigma^2 estimated as 0.1272:  log likelihood = -11.88,  aic = 29.75
```

According to the above models, we can confirm ARIMA(0,1,1) is an appropriate model since it gives the smallest AIC value.

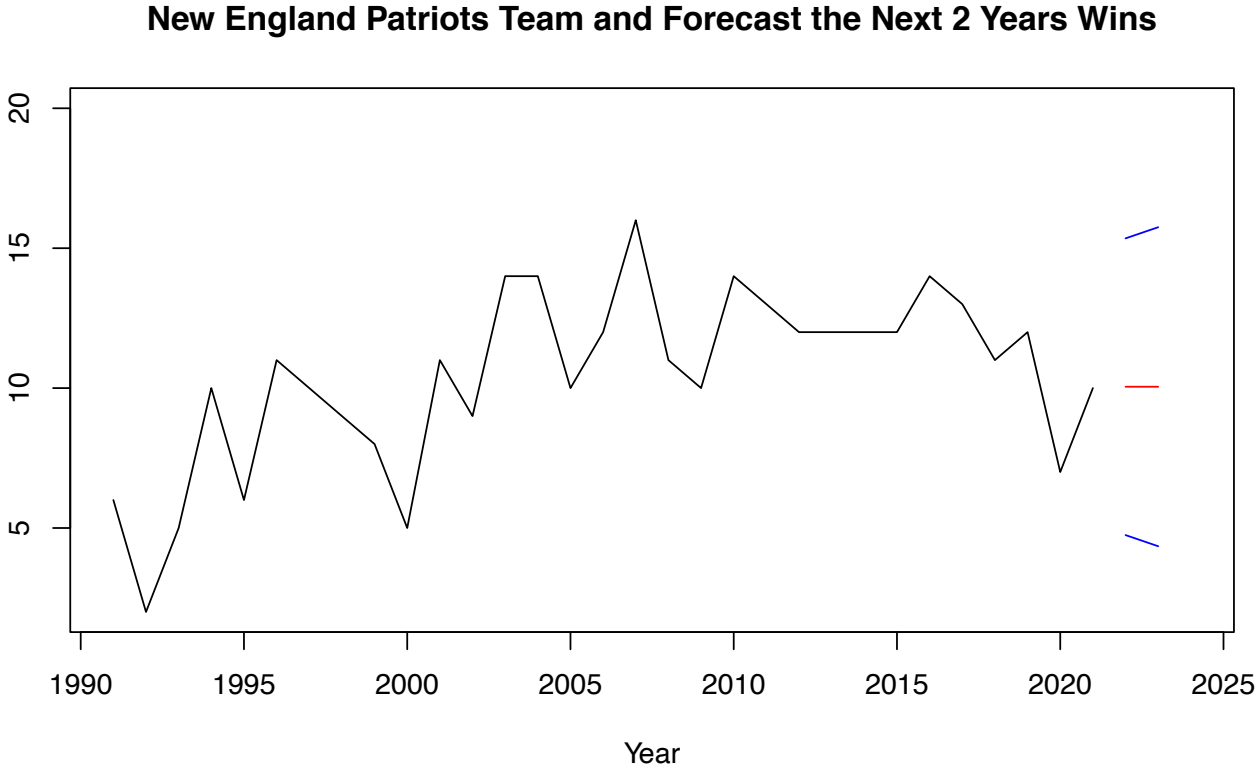
Residuals Diagnostic



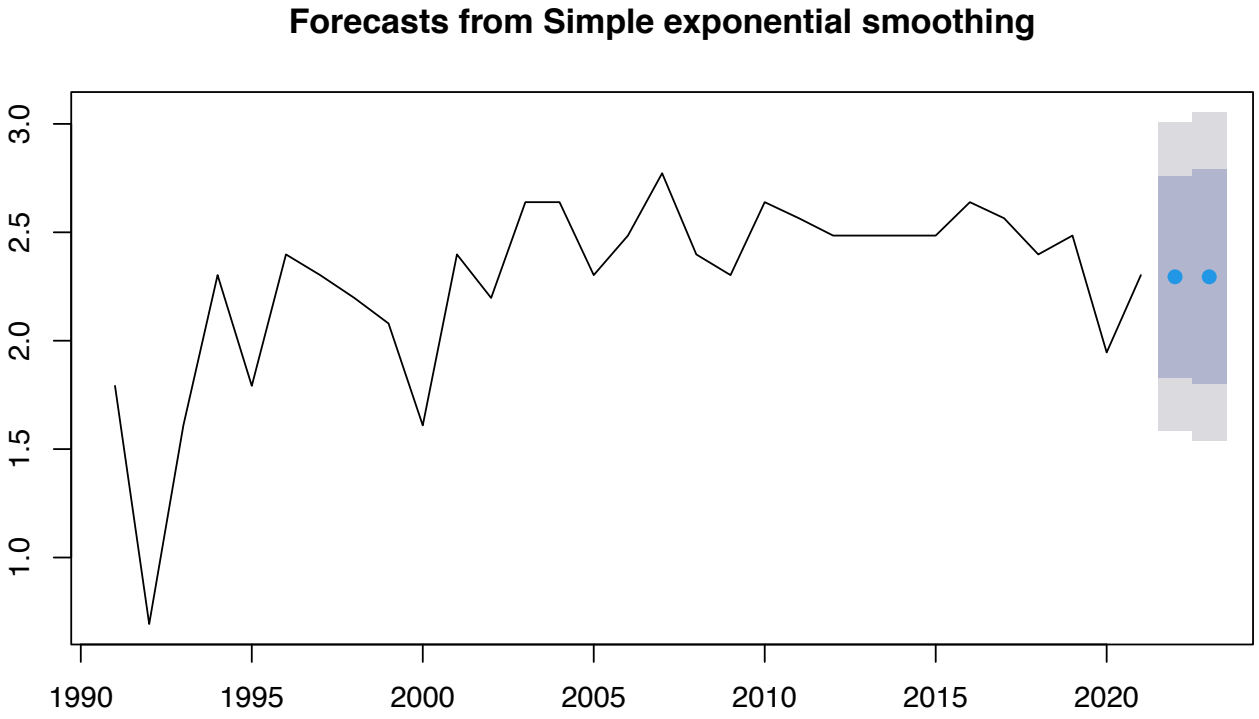


The sample ACF and PACF plot on the diagnostic suggest the residuals follow white noise, and QQ plot indicates the normality assumption doesn't seem to be violated. Overall, the fitted model looks good.

Forecasting



The plotted series above is based on the `predict()` function in R, and contains 2 arguments. In the first argument, the best fit ARIMA model, and in the second argument, the number of periods for forecasting, in this case, the following 2 years. Also, the blue lines represent the 95% forecasting limits and the red line the forecasting values for the 2 years ahead.



The visualization above, the `ses()` function in R is implemented. It stands for `simple exponential smoothing`; and the arguments `y` and `h` indicate the time series data and the number of periods for forecasting.

The simple exponential smoothing model is considered to use on stationary series. The forecasting based on this model is performing as expected and looks pretty similar to the prediction model.

Conclusion

Based on the two forecasting visualizations, we can appreciate `New England Patriots` team has a 95% predicted interval between 5 wins to 15 wins, with the most likely outcome being 10 wins for the next 2 years (2022-2023). As we can see, using different methods to test the NFL data lead us to fit the best forecasting model and finally plot it for a better understanding. In the future, analyzing a little further the team (i.e, players) would be good option to obtain a better forecasting prediction.

Limitations

Creating a prediction model can lead us to perform different models, but sometimes is difficult when it is the right time to stop while we are building the best model. Also, in the NFL dataset, we have to think about the team as a whole.

- if the team will stay the same for the next 2 years (i.e., players, coach, . . .)
- if the best players receive any injuries during the next 2 regular seasons. That can affect their overall performance.

Appendix

```
knitr::opts_chunk$set(
    echo = FALSE,
    fig.height = 5,
    fig.width = 8,
    message = FALSE,
    warning = FALSE,
    results = "hide"
)
library(tidyverse) #data structure
library(ggplot2) #visualization
library(forecast) #prediction
# import data
football <- readxl::read_xlsx("Football_Project.xlsx")
#head(football)
# create a time series object
football_ts <- ts(data = football[,5], start = 2001, end = 2021, frequency = 1)
# Plot of all teams in AFC football division
ggplot(football, aes(x = Year, y = W)) +
  geom_line(aes(color = Tm, group = Tm)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(y = "Wins") +
  facet_wrap(~ Tm)
NFL <- readxl::read_xlsx("Football_Project2.xlsx")
#NFL
NFL2 <- NFL[order(NFL$Tm),]
#NFL2


P <- ggplot(NFL2, aes(x=Year,y=W,colour=Tm,group=Tm)) + geom_line()

P2 <- P + facet_wrap(~Tm) +
  theme(strip.text.x = element_text(size = 10),
        axis.text.x = element_text(colour="grey20",
                                   size=12, angle=90, hjust=.5, vjust=.5))

P2
# filter a team from the dataset
football_team <- filter(.data = football, Tm == "New England Patriots")

# create a time series object
football_team_ts <- ts(data = football_team$W, start = 1991, end = 2021,
                       frequency = 1)

# plot of the original series
plot(football_team_ts, xlab="Year", ylab = "",
     main = "Wins of New England Patriots Team")

# plot of the logarithms of the series
football_log <- log(football_team_ts)
plot(football_log, xlab="Year", ylab = "",
     main = "Log Wins of New England Patriots Team")
```

```r
#plot of the first difference of the logged series.
diff1 <- c(NA, diff(log(football_team_ts)))
diff1 <- ts(diff1, start = c(1991,1), deltat = 1)

plot(diff1, xlab = "Year", ylab = "", main = "First order differenced series")
## ACF & PACF plots
par(mfrow = c(1, 2))
acf(football_team_ts)#, lag.max = 15)
pacf(football_team_ts)#, lag.max = 15)

auto.arima(football_team_ts)

par(mfrow = c(1, 2))
acf(diff1, lag.max = 12, na.action = na.pass, main = "ACF for differenced series")
pacf(diff1, lag.max = 12, na.action = na.pass, main = "PACF for differenced series")

auto.arima(diff1)
fit_ar1 <- arima(football_log, order = c(p = 1, d = 0, q = 0))
fit_ar1
fit_ar2 <- arima(football_log, order = c(p = 2, d = 0, q = 0))
fit_ar2
fit_arima <- arima(football_log, order = c(p = 0, d = 1, q = 1))
fit_arima
fit_arima1 <- arima(football_log, order = c(p = 1, d = 1, q = 1))
fit_arima1
## Fitting residuals
par(mfrow = c(1, 2))
res <- fit_arima$residuals

# ACF & PACF plots
acf(res, lag.max = 12, na.action = na.pass)
pacf(res, lag.max = 12, na.action = na.pass)

# diagnostic
tsdiag(fit_arima)
# check normality of the residuals
qqnorm(res)
qqline(res)
# prediction of the time series
pred <- predict(fit_arima, n.ahead = 2)

# prediction plot
plot(football_team_ts, xlim = c(1991, 2024), ylim = c(2,20),
     ylab = "", xlab = "Year",
     main = "New England Patriots Team and Forecast the Next 2 Years Wins")

## forecasted values
lines(exp(pred$pred), col = "red")

## 95% forecasting limits
lines(exp(pred$pred-2*pred$se),col='blue')
lines(exp(pred$pred+2*pred$se),col='blue')
```

```r
## Simple exponential smoothing
forecast_team <- ses(y = football_log, h = 2)
plot(forecast_team)
```