The Juiced Ball Theory

105070038

Wei-Yu Tseng

Introduction

2019 was the Year of Home Runs, the previous record for home runs in a season by the league as a whole was 6,105. This season, there were 6,776 home runs. Two teams hit more than 300 home runs, the Twins with 307 bombs and the Yankees followed up with another record-breaking 306 dingers. Before this, the record for home runs in a season by a team was 267. Moreover, 4 teams went over the line in '19 season, and 15 teams set franchise home run records, with more that will be shown later in Exploratory Data Analysis.

People who have followed the Major League Baseball games recently may already have observed the phenomenon. — The home run rate increased dramatically, but not at all for the better.

Many explanations have been put forth: The Fly-Ball Revolution, PEDs and Steroids, etc. The most commonly mentioned conspiracy is "The Juiced Ball Theory".

The theory states that the baseballs used in Major League Baseball have been deliberately altered by the league in order to increase scoring.

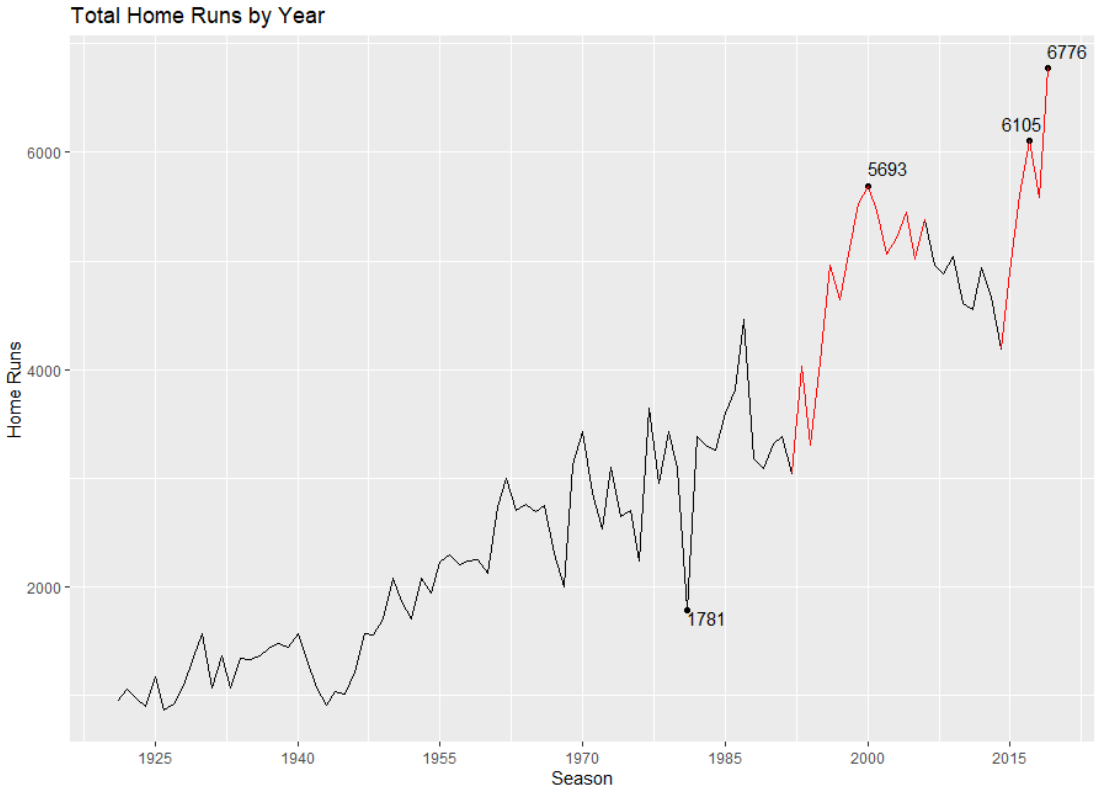
This was not the first time the "Juiced Ball Theory" was mentioned, in fact, the theory first came to prominence in late '90s to early 2000s but faded away since it became clear that the offense boom during the time was due to the usage of steroids and the PEDs. But it made a resurgence recently, thanks to the uncanny uptick in offensive output, especially the home runs.

Therefore, our goal is to detect whether the anomaly of the offensive outbreak exists using time-series analysis.

Exploratory Data Analysis

The data from 1921 to 2019 are used here, because the baseball rules were different before 1920, then named the "Dead-Ball Era".

First, take a look at the total home runs hit by year over the span.



It is easy to observe an increasing trend of home run productions with some micro fluctuations over the time, but there seem to be some eccentric patterns among the results.

Due to the strike, 713 games of the 1981 season were cancelled (38% of the MLB schedule), and only 1781 home runs were hit that year, but besides this year, there still existed some abnormal home run productions over two time periods, one was around late '90s to early 2000s, and the other was the span over the last five seasons (2015~2019).

Steroids and the PEDs

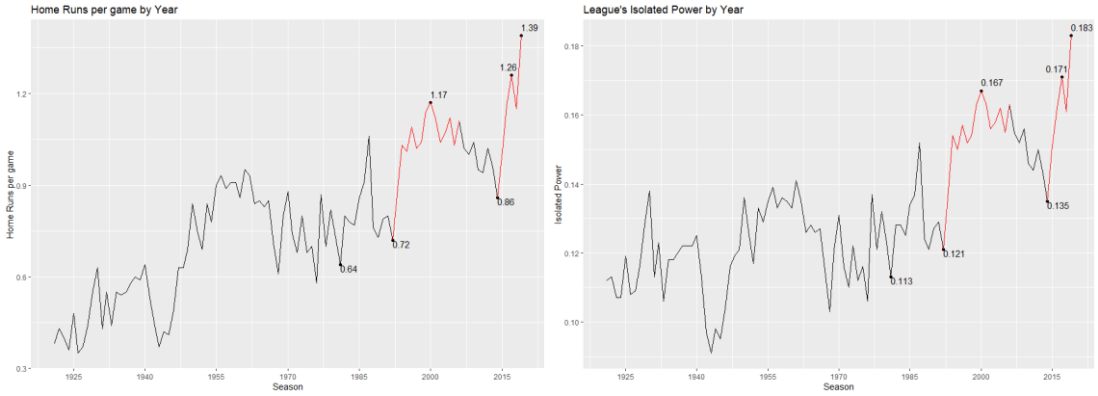
2007, the "Mitchell Report" revealed the shocking fact that numerous Major League Baseball players have used steroids and performance-enhancing drugs (PEDs) to improve their physical and mental abilities such as muscles and coordination, which explained the aberrant increase in offense over "The Steroids Era".

Experts and Scientists believe that "The Steroids Era" began at some point around the late 90s and terminated in 2007, the time "Mitchell Report" was published.

Above plot shows some clues about "The Steroids Era", where the home run rate

skyrocketed ridiculously year by year from 1990 to 2000, maintained at such level for a couple of years, and then declined after 2007.

But what happened over the last five seasons?



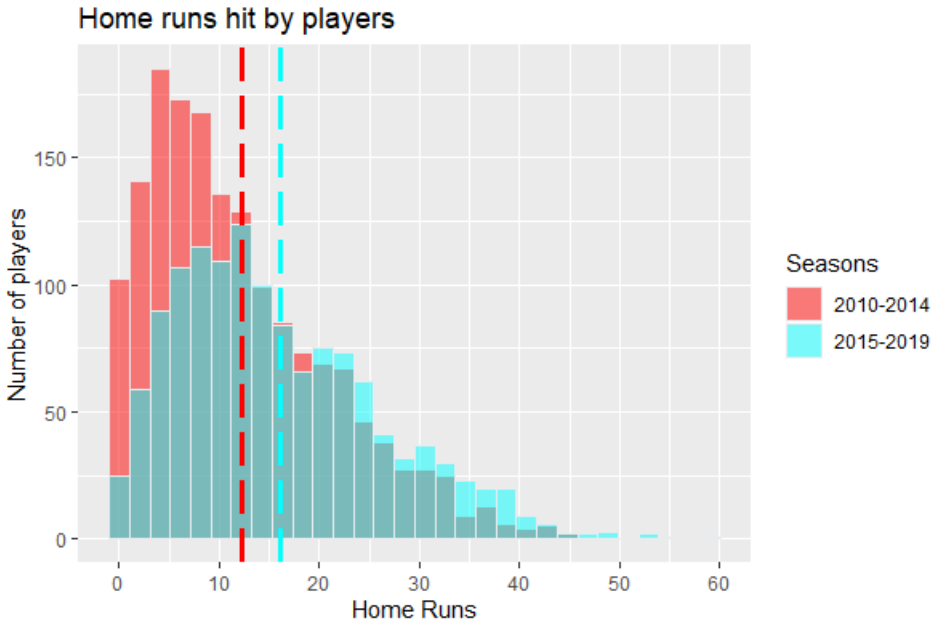
Instead of using total home runs production by year, the home run numbers per game might be a better stat to investigate here, since the numbers of games per season several decades ago were different from now.

The number of home runs went from 0.86 per game to 1.39 per game within four years, a 61.6% increase in total, and approximate 10.08% per year. Even when during the "Steroids Era", the annual increase rate was "only" 6.25% per year, barely above half of the modern years. (Annual increase rate over the century was 1.3% per year.)

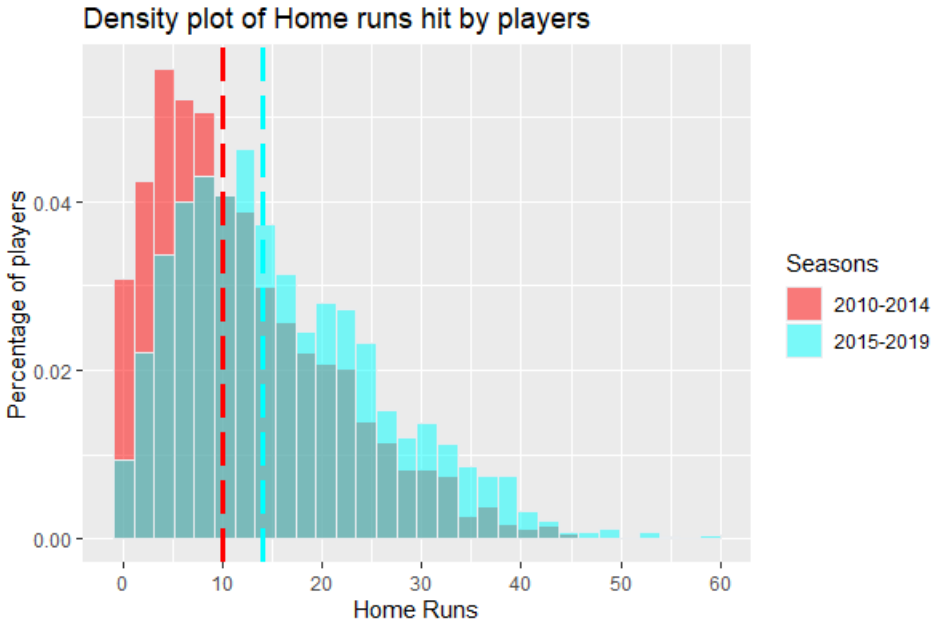
The inclination of Isolated Power also demonstrates the same thing.

We can also take a look at the difference between the home run distributions of two groups.

First, separate the data from 2010 to 2019 into two groups, 2010 to 2014 and 2015 to 2019, respectively.



Two vertical dash-lines are the mean home runs of each distribution. In the later seasons (2015-2019), the hitters seemed to hit more home runs than the early seasons (2015-2019), means and medians of both groups also imply the same thing.



The density plot again indicates that the top hitters performed clearly better in later seasons. (Two vertical dash-lines here are the median home runs of each distribution.)

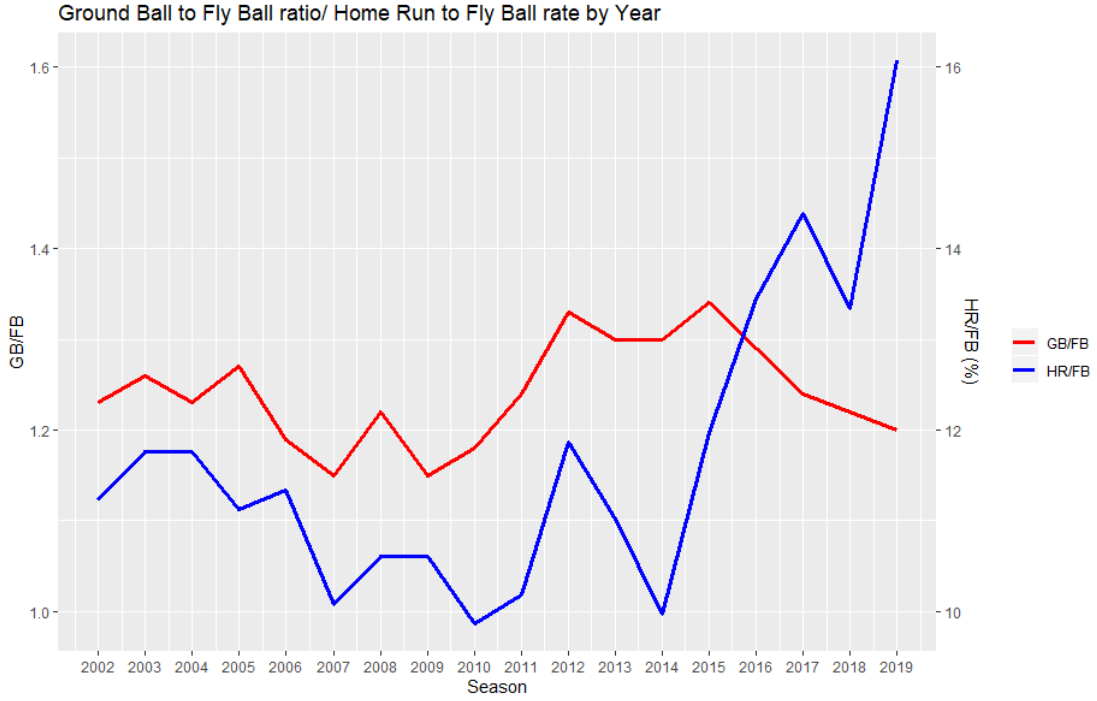
```
wilcox.test(G1014$HR,G1519$HR,alternative = "less", paired = FALSE, conf.level = 0.95)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  G1014$HR and G1519$HR
## W = 822497, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

The p-value of the Mann–Whitney test also suggests the difference between two samples' medians.

After the "Mitchell Report" was published, the MLB officials have been seriously executing the drug test, some even claimed they were tested twice within a week; the penalty and punishment of using steroids were nothing half severe — more than 50 major leaguers have been suspended over 50 games, the two famous cases were Alex Rodriguez's 162 games in 2013 and Jenrry Mejía's permanent suspension. Under such regularities and punishment, the players are not supposed to take the risks. Therefore, the most commonly seen explanations are the "Fly-Ball Revolution" and the "Juiced Ball Theory".

The Fly-Ball Revolution

The defensive shifts have become more and more common since the success of Oakland A's moneyball theory in 2002, hitting ground balls has been the less effective way to get on base. Studies show that hitting fly balls generates more offensive values than ground balls, and since home runs are fly balls (except for very rare inside-the-park home run cases), it's easy to connect to the methodology that hitting more fly balls yields more home runs.



From the above plot, a slightly decreasing trend of "Ground Ball to Fly Ball ratio (GB/FB)" over the last five seasons (2015-2019) can be observed, the same time horizon as the outburst of home runs, but it is not some particularly bizarre pattern. As we can see, the similar pattern could be observed as well in the early seasons, so this may only be some random but logical fluctuations. Moreover, even though the fly ball rate increased, it shouldn't affect the corresponding "Home Run to Fly Ball rate (HR/FB)", that is, the HR/FB may also be alike to a series with its own trend and pattern instead of the eccentric uptick in the graph. Therefore, this assumption may be put behind of the "Juiced Ball" conspiracy.

The Juiced Ball Theory

The similar argument in the previous paragraph can explain the "Juiced Ball Theory", especially the abnormal rise in HR/FB.

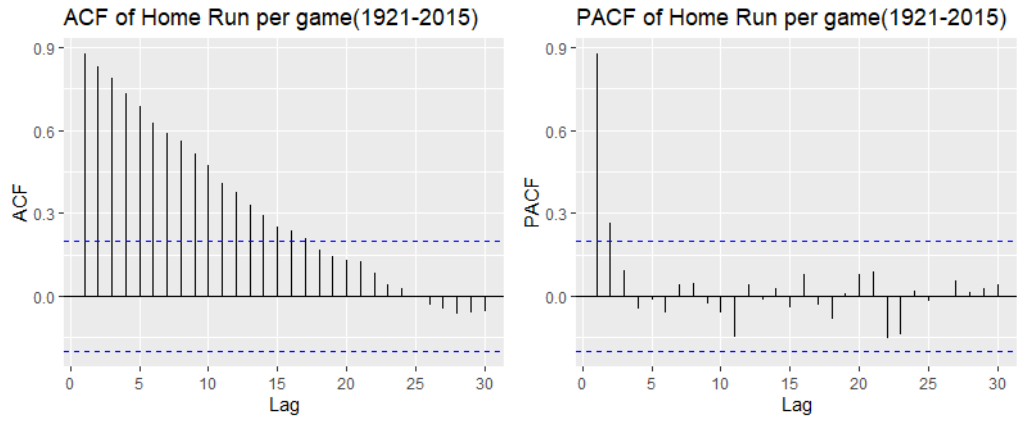
Model Building

Two Datasets "Home Runs per game by year" and "Isolated Power by year" will be used to perform the following analysis.

Home Runs per game by year

First separate the data into two parts, training data(1921-2015) and testing data(2016-2019).

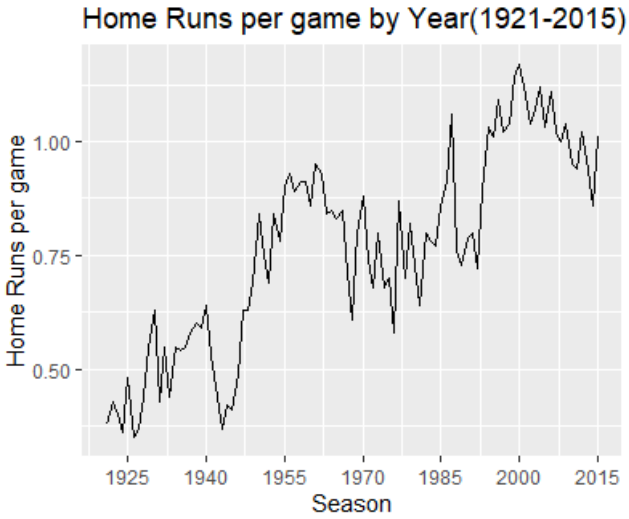
Acf and Pacf



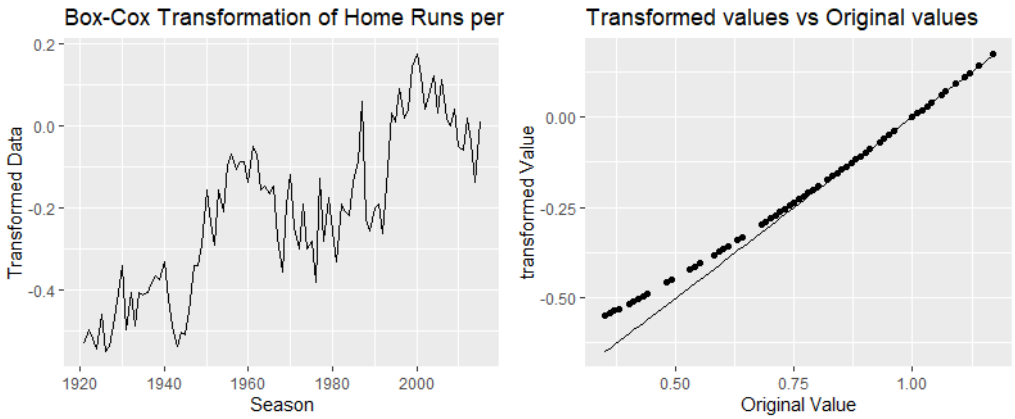
The ACF of the training data tails off, while the PACF cuts off after lag 2 suggesting that an AR(1) or even a weak AR(2)structure exists among the series without any MA term.

Transformation and Differencing

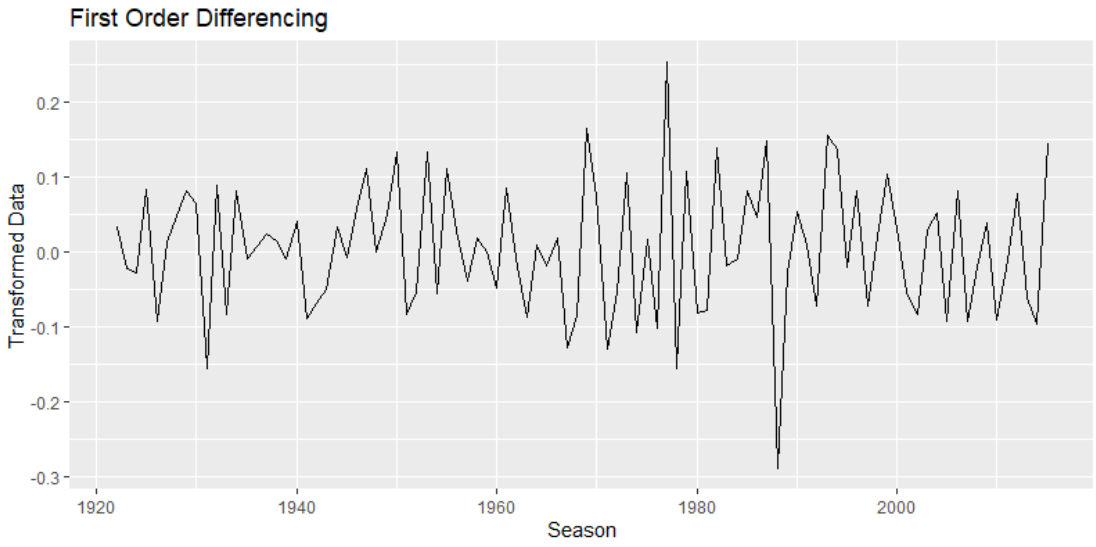
In order to perform ARMA modeling, the series must first be stationary. That is, the heteroscedasticity should be replaced by homoscedasticity, and the mean structure also needs to be eliminated.



The series seems to be gaussian-like already, but to prevent any extra error, the Box-Cox transformation will still be applied here.



The transformed dataset and the original dataset are very much alike, λ here is 1.398 which is very close to 1.



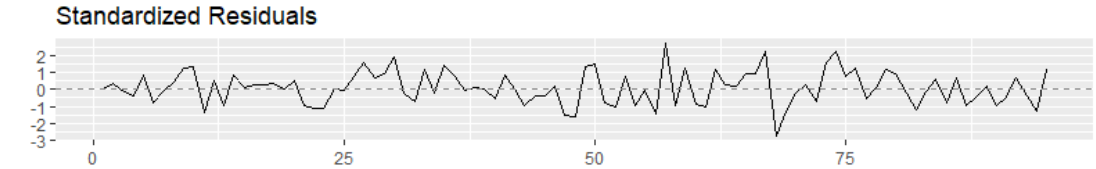
The mean structure seems to be eliminated after the first-order differencing, and the series looks much more stationary then the original data.

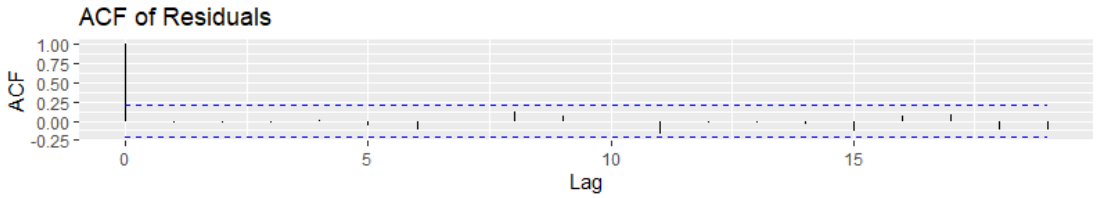
Therefore, the ARIMA(2,1,0) is our first experiment suggested by the above characteristics of the series.

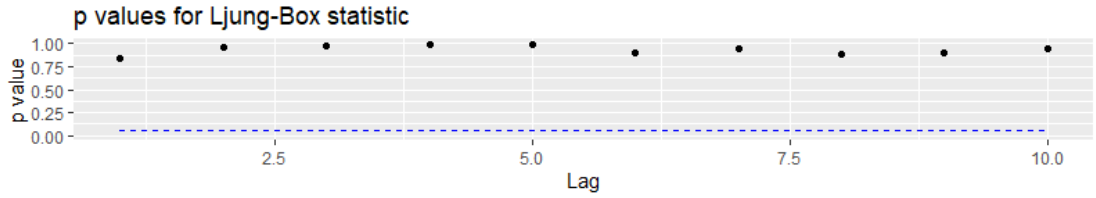
Modeling

ARIMA(2,1,0)

```
arima(x = y, order = c(2, 1, 0))

Coefficients:
          ar1      ar2
      -0.3840  -0.2047
s.e.   0.1017   0.1016

sigma^2 estimated as 0.006561:  log likelihood = 102.77,  aic = -199.55
```



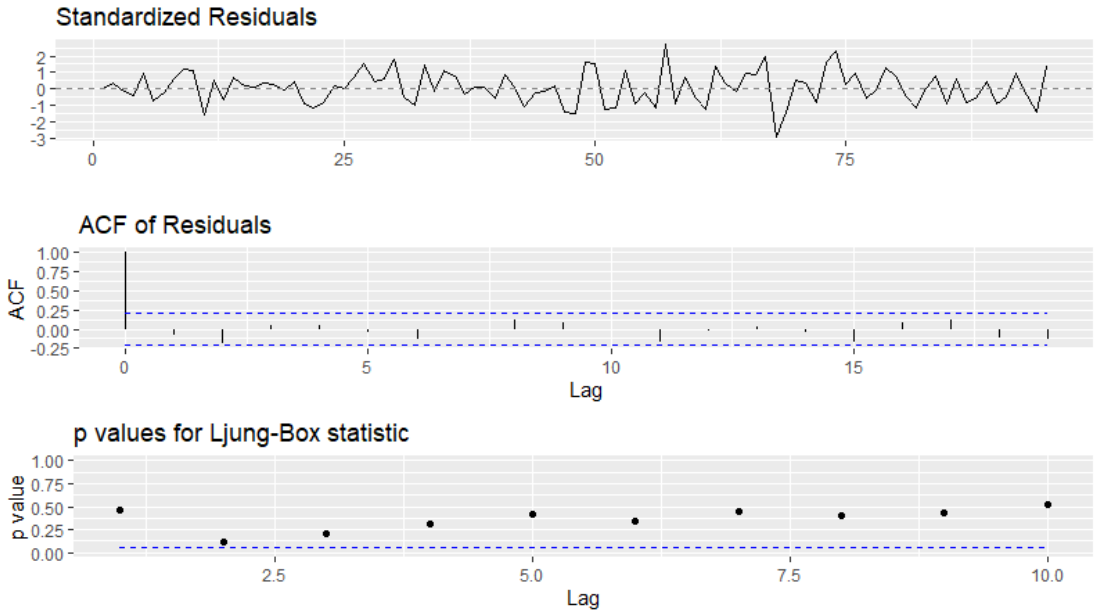




Although the diagnostics here look decent for the ARIMA(2,1,0) model, the AR2 here is on the borderline (the 95% C.I. almost covers 0), thus, dropping AR2 and reconsidering about ARIMA(1,1,0) as another option.

ARIMA(1,1,0)

```
arima(x = y, order = c(1, 1, 0))

Coefficients:
          ar1
      -0.3212
s.e.   0.0987

sigma^2 estimated as 0.00685:  log likelihood = 100.79,  aic = -197.58
```

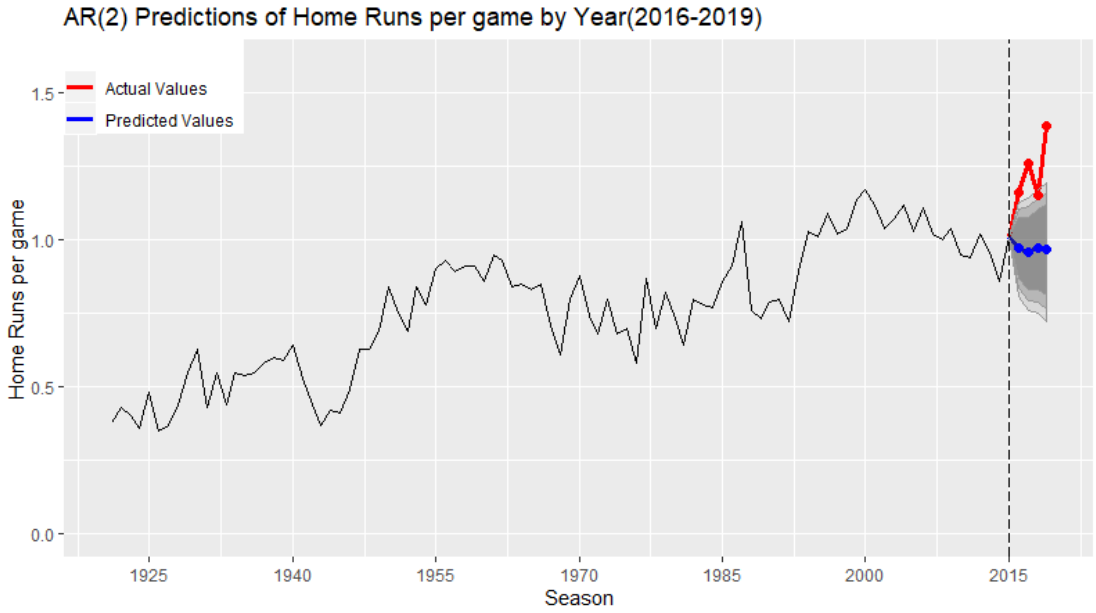


The results are still pretty good, but both LJB test result and AIC suggest that ARIMA(2,1,0) should be our final selection over ARIMA(1,1,0), but the further analysis will give a totally different inspiration.

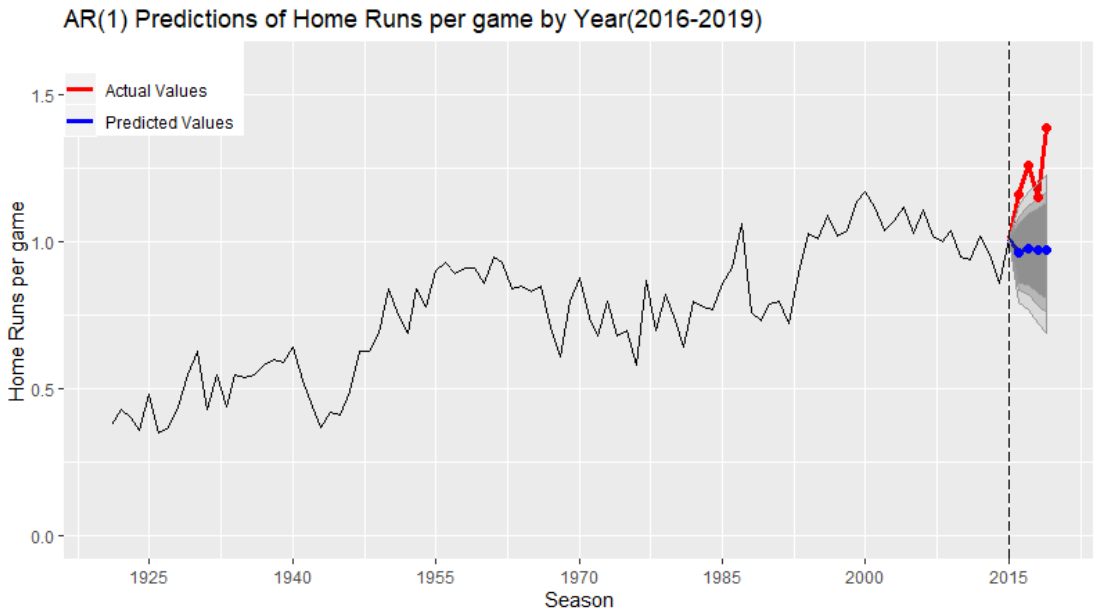
Forecasting

Since the final model selection has not been determined, both models will be used to forecast the home runs production per game for '16 - '19 seasons.

ARIMA(2,1,0)



ARIMA(1,1,0)



The 3 Polygons with different color depth represent 80%, 90% and 95% confidence regions respectively.

Predictions of ARIMA(2,1,0) are slightly more conservative than ARIMA(1,1,0) with lower predicted values and narrower confidence region, but there isn't much difference.

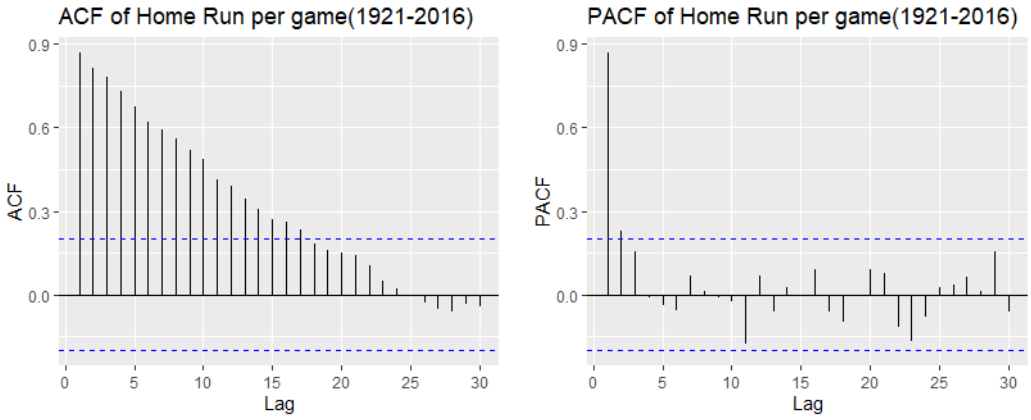
Moreover, 3 of the 4 actual values lie outside of the 95% confidence region of both models, the other one lies outside of the 90% confidence region.

Iteration

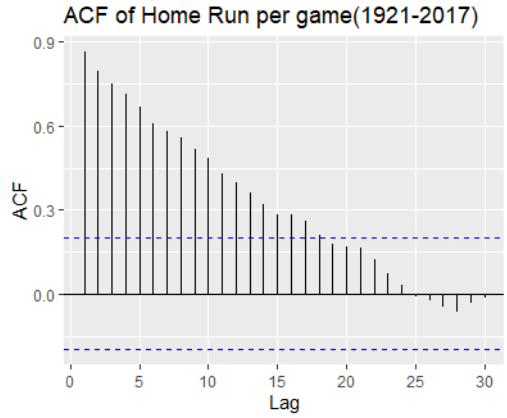
Repeat the above steps using 1921-2016, 1921-2017 and 1921-2018 as training dataset, 2017-2019,2018-2019 and 2019 as testing dataset to construct 3 different models with analogous method.

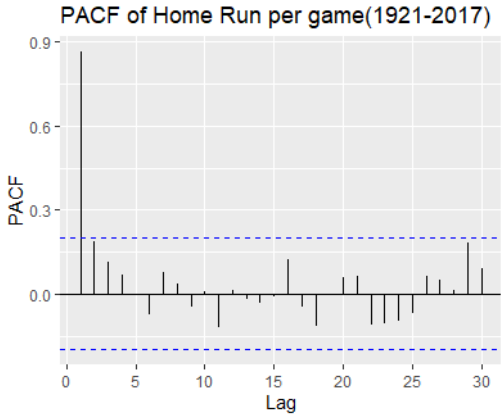
Acf and Pacf

1921-2016

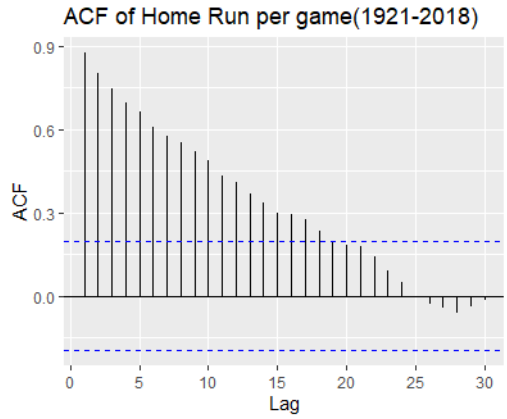


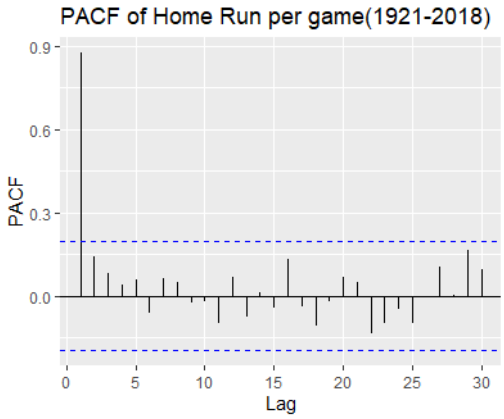
1921-2017





1921-2018





Modeling

1921-2016

ARIMA(2,1,0)

```
arima(x = y2, order = c(2, 1, 0))

Coefficients:
          ar1      ar2
      -0.3530  -0.1568
s.e.   0.1028   0.1032

sigma^2 estimated as 0.01161:  log likelihood = 76.79,  aic = -147.57
```

ARIMA(1,1,0)

```
arima(x = y2, order = c(1, 1, 0))

Coefficients:
          ar1
      -0.3021
s.e.   0.0983

sigma^2 estimated as 0.0119:  log likelihood = 75.65,  aic = -147.29
```

1921-2017

ARIMA(2,1,0)

```
Coefficients:
          ar1      ar2
      -0.2936  -0.1657
s.e.   0.1020   0.1030

sigma^2 estimated as 0.007092:  log likelihood = 101.26,  aic = -196.52
```

ARIMA(1,1,0)

```
arima(x = y3, order = c(1, 1, 0))

Coefficients:
          ar1
      -0.2480
s.e.   0.0992

sigma^2 estimated as 0.007288:  log likelihood = 99.98,  aic = -195.97
```

1921-2018

ARIMA(2,1,0)

```
Coefficients:
          ar1      ar2
      -0.3303  -0.1372
s.e.   0.1005   0.1003

sigma^2 estimated as 0.01142:  log likelihood = 79.23,  aic = -152.46
```

ARIMA(1,1,0)

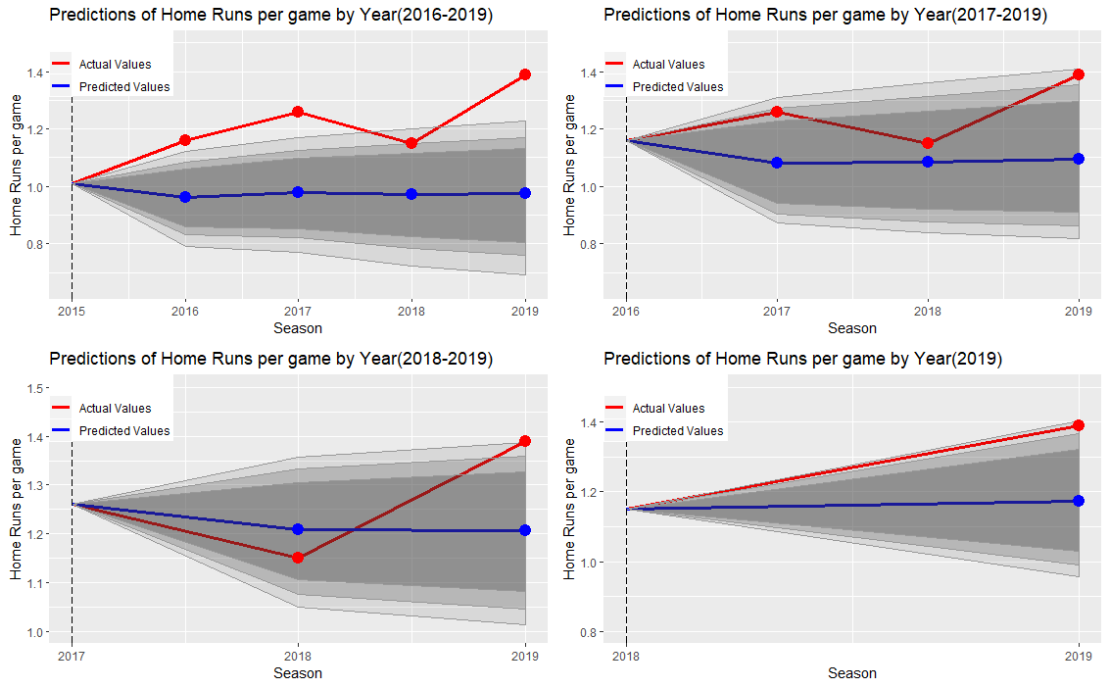
```
Coefficients:
          ar1
      -0.2910
s.e.   0.0972

sigma^2 estimated as 0.01164:  log likelihood = 78.3,  aic = -152.61
```

Neither of the coefficients of AR2 in 3 models is significant, with likelihood, AIC and estimated sigma square are only slightly better(smaller) in ARIMA(2,1,0), which introduces the possibility of overfitting. Since the universal model of home runs per game should be identical, the ARIMA(1,1,0) is our final selection for further analysis and forecasting, with the general fitted model being closed to the form ：

$$(\mathbf{1}+\mathbf{0.3}\boldsymbol{B})\boldsymbol{X_t}=\boldsymbol{\epsilon_t},\boldsymbol{\epsilon_t}\sim\boldsymbol{N}(\mathbf{0},\mathbf{0.011}),$$

where $X_t = Home\ Runs\ per\ game\ at\ season\ t - Home\ Runs\ per\ game\ at\ season\ t-1$.

Predicted values vs Real values (testing data)



Most of the actual values didn't lie within the 90% confidence region, or even the 95% confidence region, even when there was a cutback in 2018, it only lay within the 80% confidence region of Model 2 and Model 3.

In the first model, none of the actual values lay within the 90% confidence region, and even only 2018 lay within the 95% C.I.; in the second model, 2018 lay within the 80% C.I. while others were outside of 90% C.I.; in the third model, 2018 was inside the 80% C.I. wihile 2019 was outside of 95% C.I.; in the fourth model, 2019 was outside of the 90% C.I..

Summary chart of the predictions:

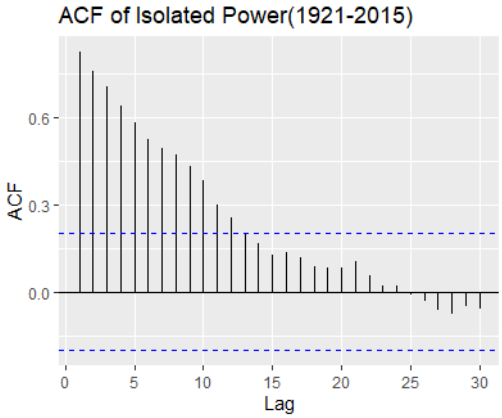
	2016	2017	2018	2019
Model 1	Outside of 95%	Outside of 95%	Outside of 90%	Outside of 95%
Model 2	NA	Outside of 90%	Inside the 80%	Outside of 90%
Model 3	NA	NA	Inside the 80%	Outside of 95%
Model 4	NA	NA	NA	Outside of 90%

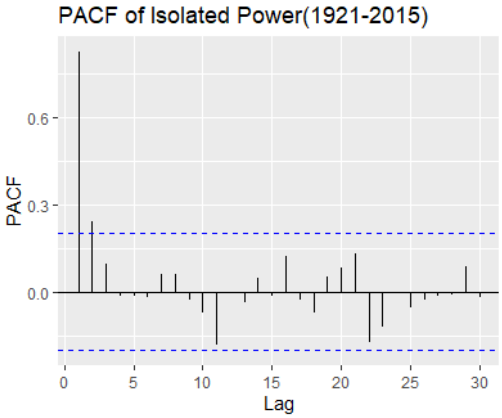
Isolated Power

Similar methods in previous experiment are applied to the data of Isolated power.

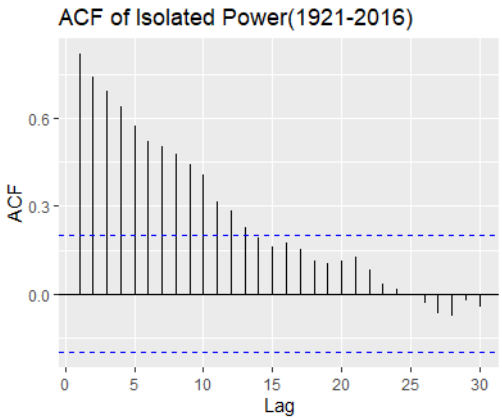
Acf and Pacf

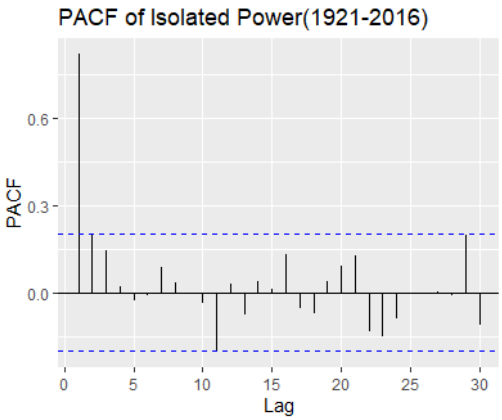
1921-2015



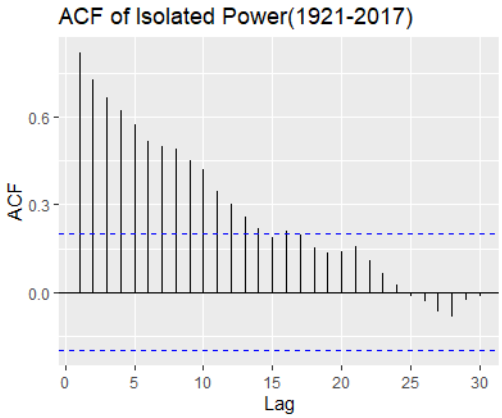


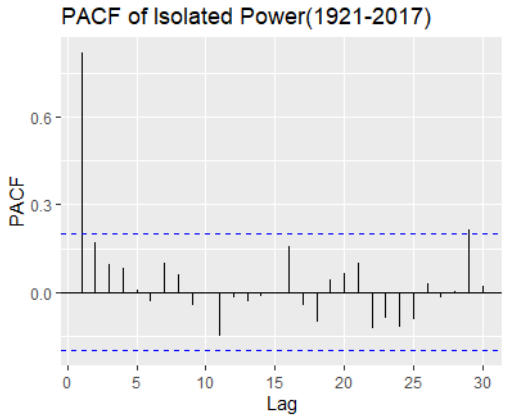
1921-2016



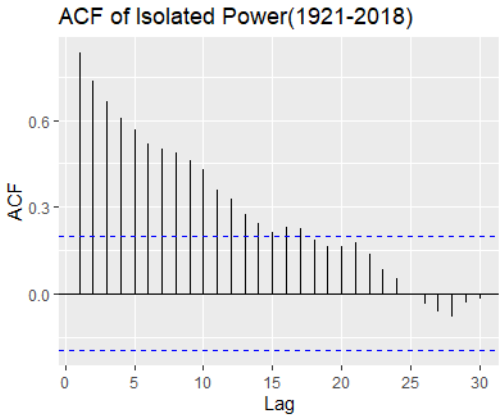


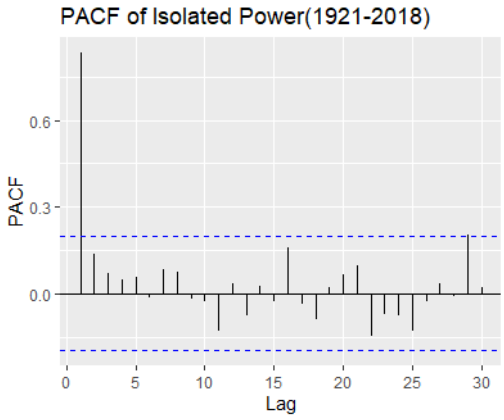
1921-2017





1921-2018





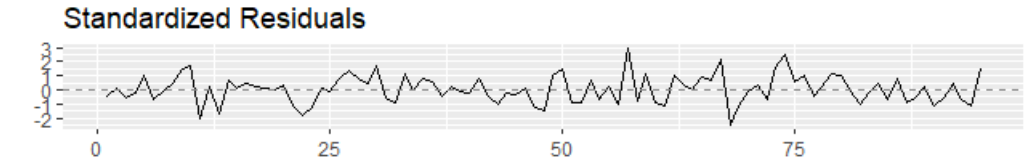
The Acfs and Pacfs of "Home Runs per game by year" and "Isolated Power" look very much alike, indicating that the model of "Isolated Power" might at least have an AR(1) structure, or even with a weak AR(2).

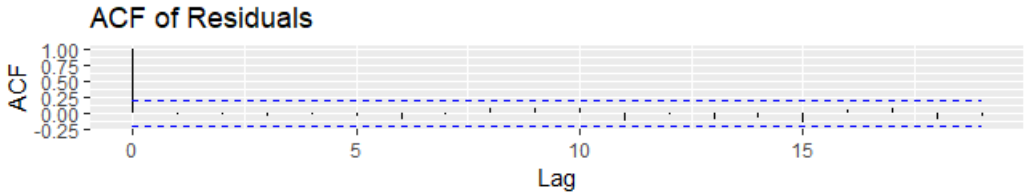
Modeling

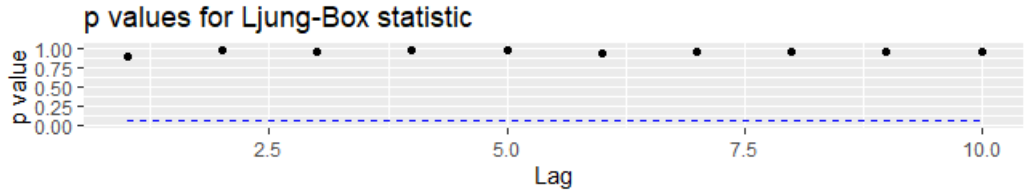
1921-2015

ARIMA(2,1,0)

```
Coefficients:
          ar1      ar2
      -0.3589  -0.1828
s.e.   0.1024   0.1023

sigma^2 estimated as 1.406e-06:  log likelihood = 499.84,  aic = -993.68
```

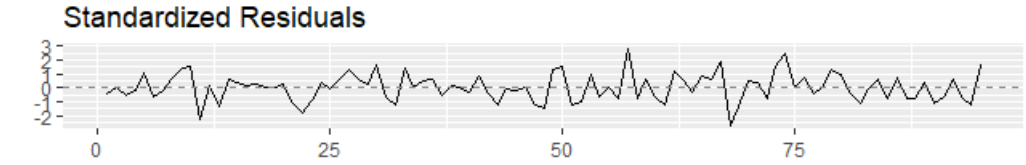


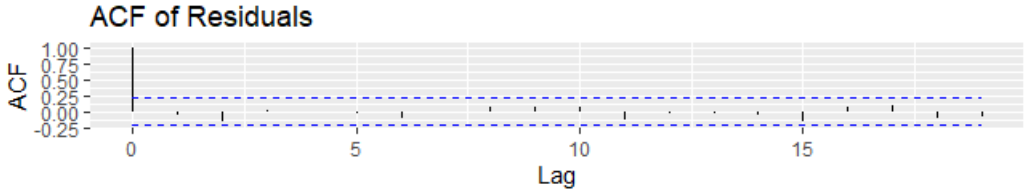


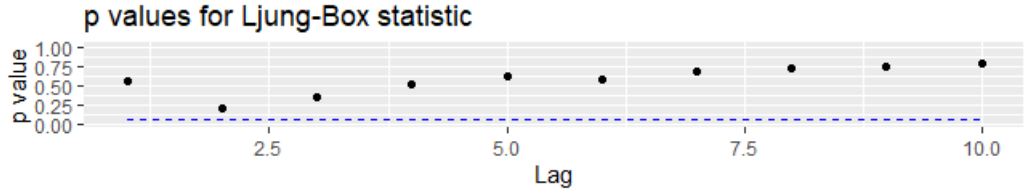


ARIMA(1,1,0)

```
Coefficients:
          ar1
      -0.3052
s.e.   0.0994

sigma^2 estimated as 1.455e-06:  log likelihood = 498.27,  aic = -992.54
```



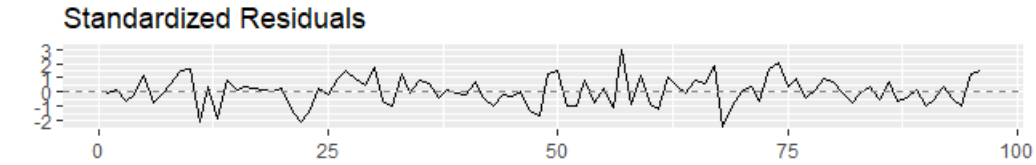


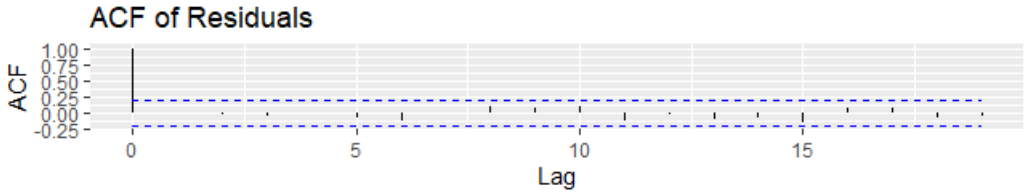


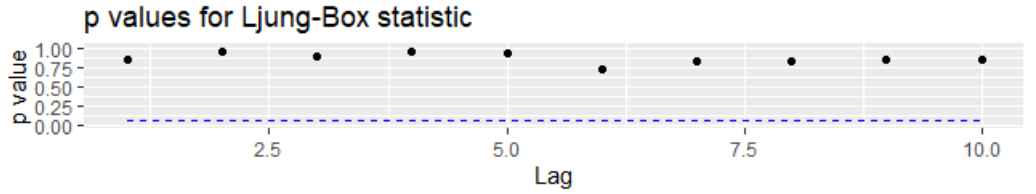
1921-2016

ARIMA(2,1,0)

```
Coefficients:
          ar1      ar2
      -0.3213  -0.1709
s.e.   0.1021   0.1030

sigma^2 estimated as 0.0001166:  log likelihood = 295.32,  aic = -584.63
```

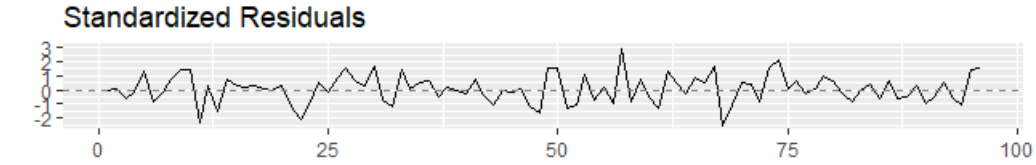


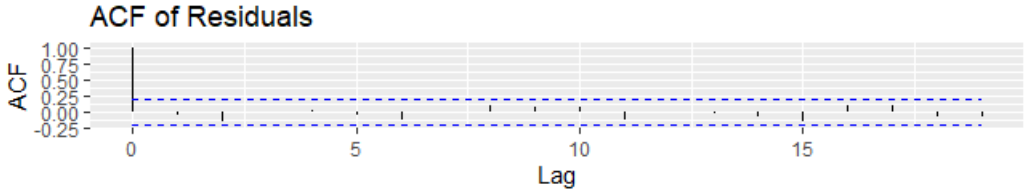


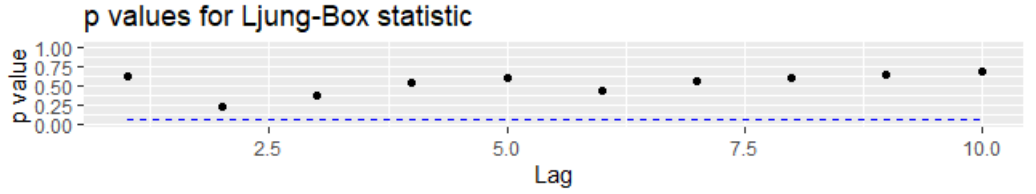


ARIMA(1,1,0)

```
Coefficients:
          ar1
      -0.2710
s.e.   0.0988

sigma^2 estimated as 0.0001201:  log likelihood = 293.96,  aic = -583.92
```



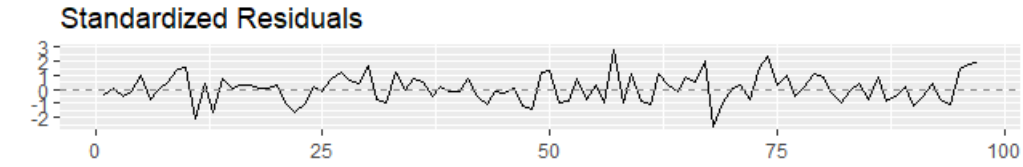


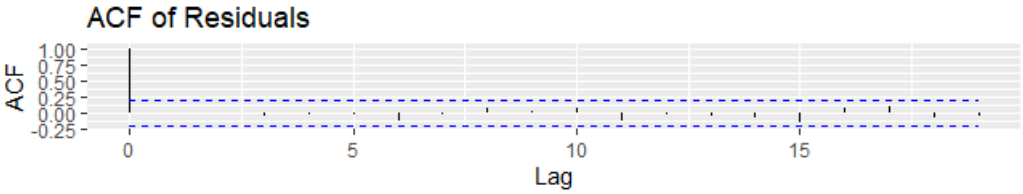


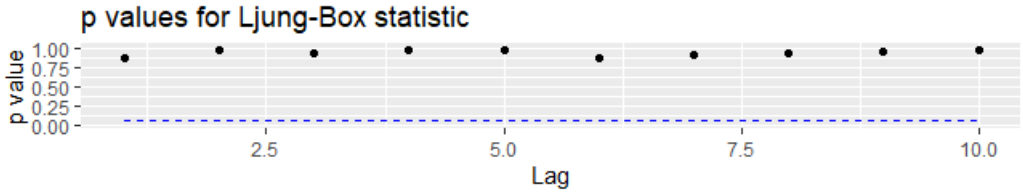
1921-2017

ARIMA(2,1,0)

```
Coefficients:
          ar1      ar2
      -0.2851  -0.1418
s.e.   0.1019   0.1024

sigma^2 estimated as 1.493e-06:  log likelihood = 507.64,  aic = -1009.27
```

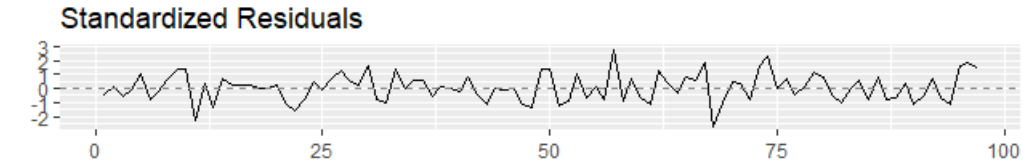


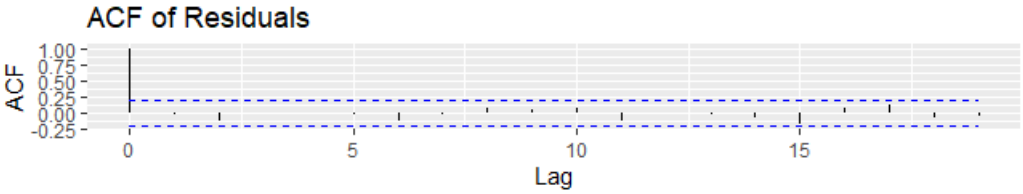


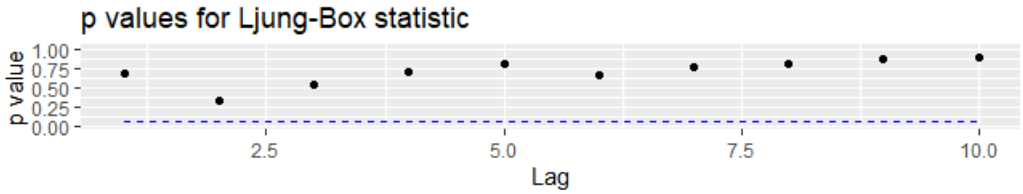


ARIMA(1,1,0)

```
Coefficients:
          ar1
      -0.247
s.e.   0.099

sigma^2 estimated as 1.524e-06:  log likelihood = 506.69,  aic = -1009.37
```



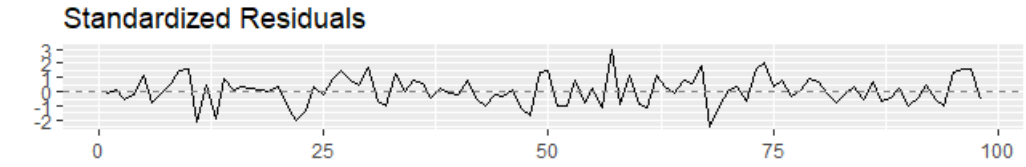


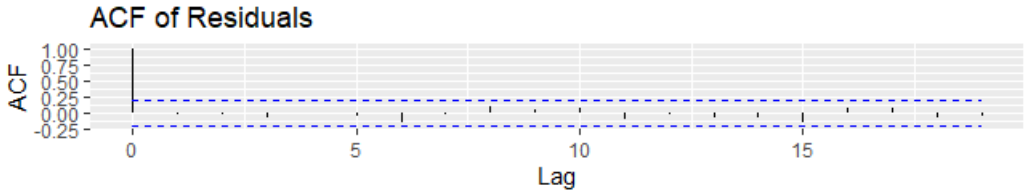


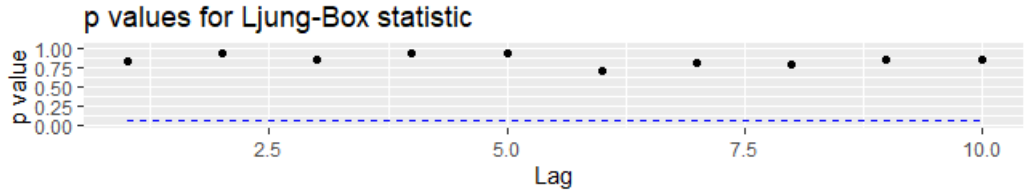
1921-2018

ARIMA(2,1,0)

```
Coefficients:
          ar1      ar2
      -0.3028  -0.1476
s.e.   0.1003   0.1003

sigma^2 estimated as 9.141e-05:  log likelihood = 313.36,  aic = -620.73
```

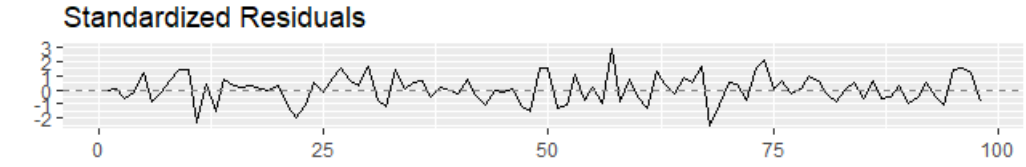


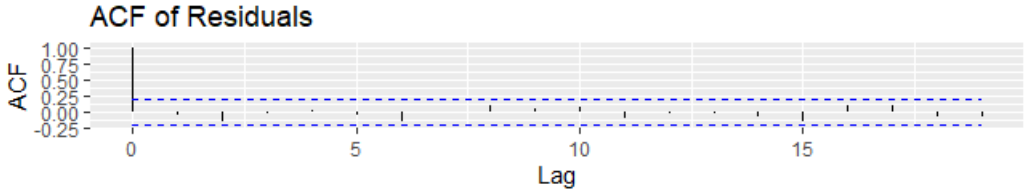


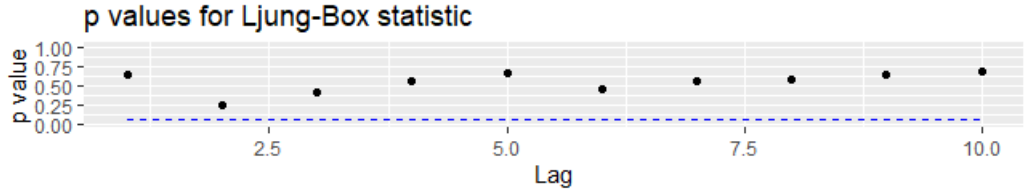


ARIMA(1,1,0)

```
Coefficients:
          ar1
      -0.2644
s.e.   0.0978

sigma^2 estimated as 9.349e-05:  log likelihood = 312.29,  aic = -620.58
```





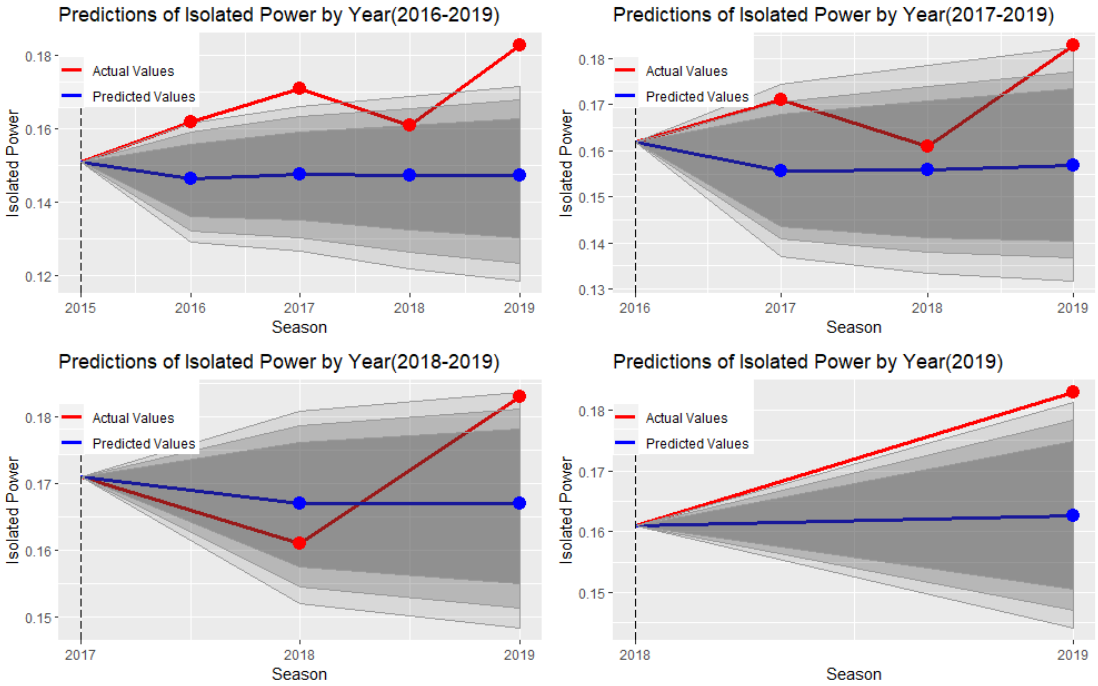


Neither of the coefficients of AR2 in 4 models is significant, with likelihood, AIC and estimated sigma square are only slightly better(smaller) in ARIMA(2,1,0), thus, the ARIMA(1,1,0) is our final selection for further analysis and forecasting, with the general fitted model being closed to the form ：

$$(\mathbf{1}+\mathbf{0.27}\boldsymbol{B})\boldsymbol{X_t}=\boldsymbol{\epsilon_t},\boldsymbol{\epsilon_t}\sim\boldsymbol{N}(\mathbf{0},\mathbf{9}\times\mathbf{10^{-5}}),$$

where $X_t = Isolated\ Power\ at\ season\ t - Isolated\ Power\ at\ season\ t-1$.

Predicted values vs Real values (testing data)



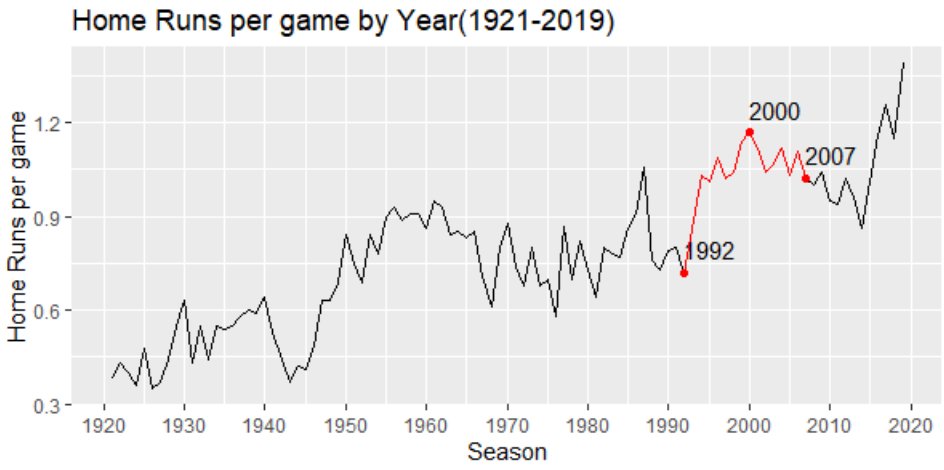
Like the results of previous experiment, most of the real values lay outside of the confidence region, with only the cutback in 2018 lying within the 80% confidence region of those models.

Summary chart of the predictions:

	2016	2017	2018	2019
Model 1	Outside of 95%	Outside of 95%	Outside of 80%	Outside of 95%
Model 2	NA	Outside of 90%	Inside the 80%	Outside of 95%
Model 3	NA	NA	Inside the 80%	Outside of 90%
Model 4	NA	NA	NA	Outside of 95%

Intervention Effect

From the results shown in the time-series modeling, several conclusions can be made, but since there were an overused quantity of PEDs and steroids 2 decades ago in baseball, we would like to know whether it affected our predictions and the real values.



As experts suggested that the prevalence of doping in baseball was around the late '90s to the early 2000s, with the aberrant offensive outburst shown in graphs (red line), the experiment will be made upon the period from 1992 to 2007, when the Mitchell Report was published.

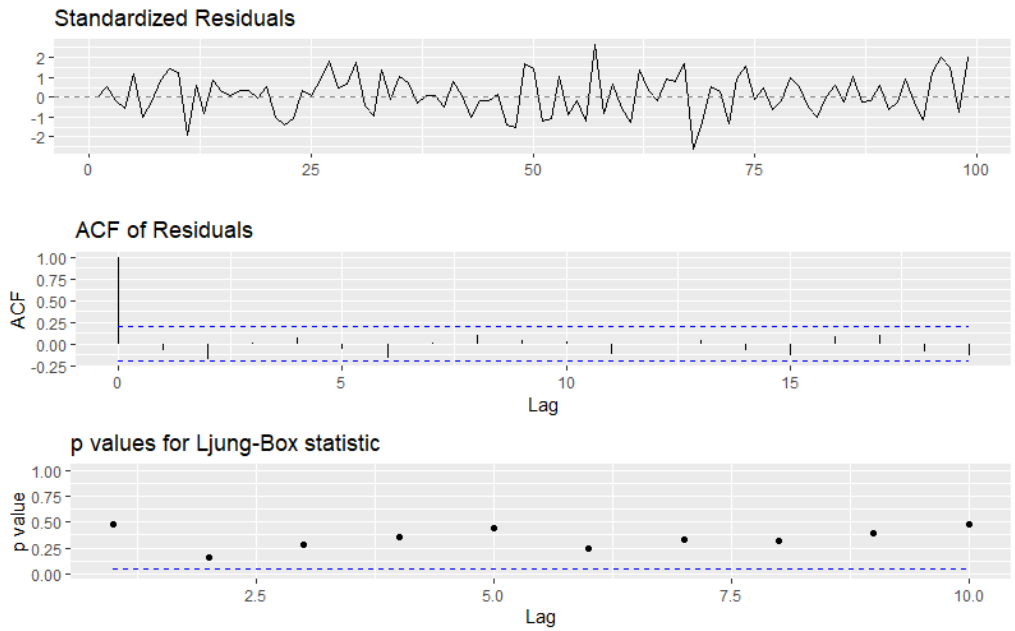
Modeling

Our results in univariate time-series analysis suggested us an ARIMA(1,1,0) model is the universal model of our two datasets, dataset "Home Runs per game" is used to show the intervention effect.

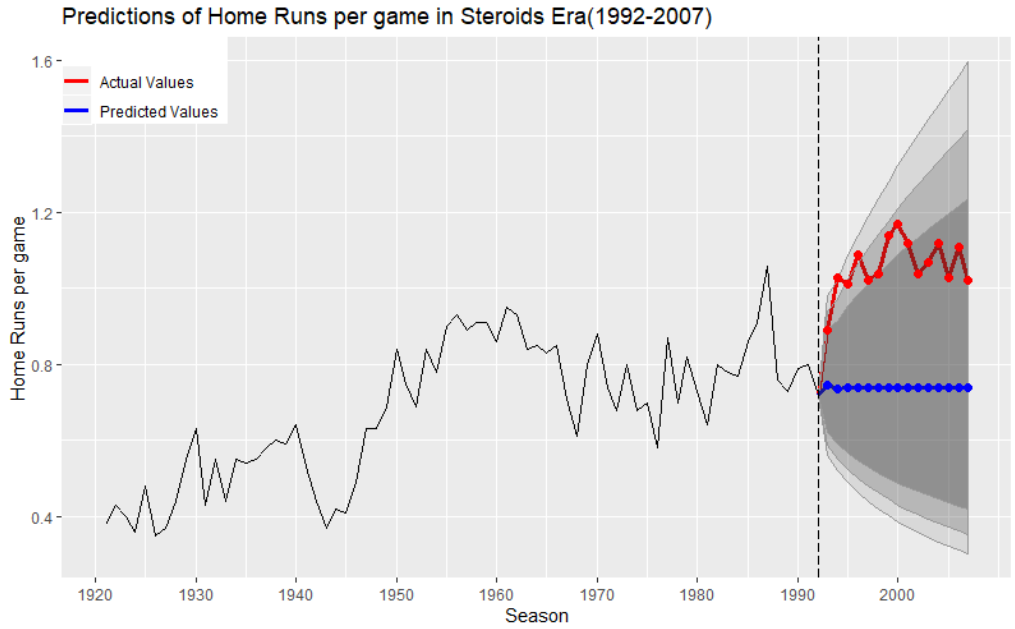
Home Runs per game

```
Coefficients:
          ar1  s9207-AR1  s9207-MA0
      -0.3183      0.761     0.0529
s.e.   0.0981      0.175     0.0399

sigma^2 estimated as 0.009704:  log likelihood = 88.02,  aic = -170.03
```

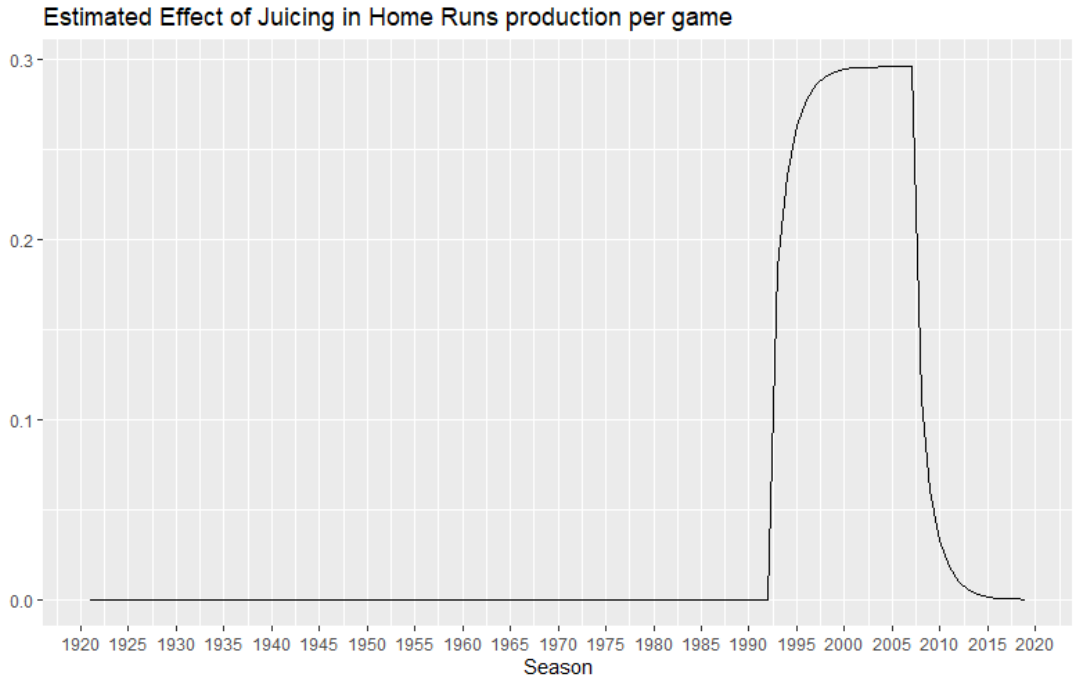
The overall AR1(-0.31) is similar to the model without transfer function, while the transfer model has a significant AR1 at 0.761 and an insignificant mean structure.



The several diagnostics of this model also looked great, no specific autocorrelation structure was detected among the residuals.



The data collected before the Steroids Era was used to construct the model for the predictions of the home runs production over the period. As shown in plot, the real values first lay outside of the 90% confidence region, but then fell back within the 80% C.I., indicating that the intervention effect may have worn off eventually.



The estimated effect of doping in baseball suggested that the intervention effect had almost worn off after the 2010 season. That is to say, the overused quantity of steroids may not have affected our model for the univariate time-series analysis in the previous paragraph.

Conclusion

Lately, people have seen many dramatical changes in Major League Baseball, from the pitcher timer (shorten the game time, reduce the pitcher's ability to change the tempo), three-batter minimum regulation (eliminate the existence of left-handed specialist) to the incoming universal designated hitter rule (designated hitters take over pitchers in batting order). These alterations somehow reveal the MLB officials' intention of increasing scoring. The Juiced Ball Theory may be one of them, stealthily,

Our experiments demonstrate the unlikeliness of the offensive bloom being the part of a coherent and rational oscillation, with most of the testing data, i.e. the real values lying outside of the 90% or even 95% confidence region, and is even more eccentric than the uptick in the Steroids Era.

Although we are still unable to attribute the abnormal home runs production to the possible alteration to baseballs, the associated skyrocketing home run to fly ball rate makes it hard for people to believe that the balls haven't been altered.

The Fly Ball revolution may also take part in the changes of offense, but comparing to these juiced balls, it is just a drop in the bucket.