Module 6: Assignment 1 - Noll Scully

Christopher Dillard

2023-06-26

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(Lahman)

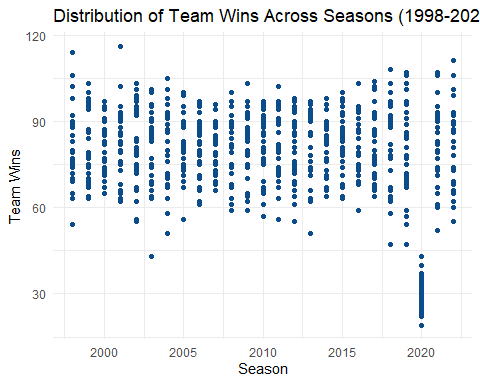
## Warning: package 'Lahman' was built under R version 4.3.1

teams = Lahman::Teams

teams = teams %>%  
 filter(yearID >= 1998)

###Task 1 Create a plot to examine how the distribution of team wins (the “W” variable) have changed over the years in the dataset. What does your plot imply about competitive balance in baseball and how that balance may change from year to year?

library(esquisse)  
#esquisser()  
ggplot(teams) +  
 aes(x = yearID, y = W) +  
 geom\_point(shape = "circle", size = 1.5, colour = "#0C4C8A") +  
 labs(  
 x = "Season",  
 y = "Team Wins",  
 title = "Distribution of Team Wins Across Seasons (1998-2022)"  
 ) +   
 theme\_minimal()



From this graph depicting annual distributions of team wins, we can see that the data points appear to constrict and expand roughly every 7 years, with varying degrees of dispersion. The anomaly seen here (where all team wins drop significantly) corresponds to the shortened 2020 season, where less games being played would logically yield a much lower mean number of wins.

###Task 2 In order to calculate the Noll-Scully metric for each year in the dataset, we need to extract a few pieces of data: 1) Mean number of wins per team. As each team plays 162 games in a regular season, we can assume that this mean value is 81. 2) Number of games played per team (we’ll use the regular season length of 162 games for this) Create an object called “ISD”. It doesn’t need to be a data frame as it will only be a single value. What is the ISD value for Major League Baseball?

ISD = 81/sqrt(162)  
ISD

## [1] 6.363961

The Idealized Standard Deviation for Major League Baseball is 6.36.

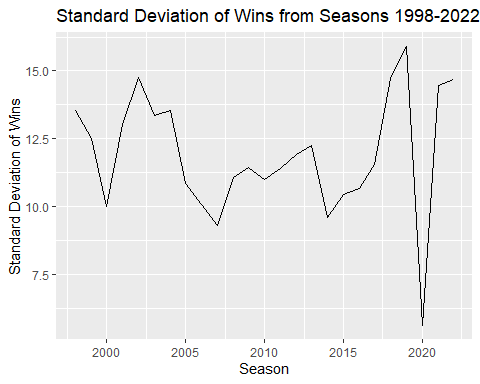
### Task 3

Use the “group\_by” and “summarize” functions to determine the standard deviation of wins (W) for each year in the dataset. How has the standard deviation of wins changed over time? Create a (nice looking) table AND chart to show this.

StddevWins\_graph = teams %>%   
 group\_by(yearID) %>%  
 summarize(StdevWins = sd(W))  
StddevWins\_graph

## # A tibble: 25 × 2  
## yearID StdevWins  
## <int> <dbl>  
## 1 1998 13.5   
## 2 1999 12.5   
## 3 2000 9.99  
## 4 2001 13.0   
## 5 2002 14.8   
## 6 2003 13.4   
## 7 2004 13.5   
## 8 2005 10.8   
## 9 2006 10.1   
## 10 2007 9.29  
## # ℹ 15 more rows

ggplot(StddevWins\_graph,aes(yearID,StdevWins))+  
 geom\_line() +  
 labs(x="Season",y="Standard Deviation of Wins",title="Standard Deviation of Wins from Seasons 1998-2022")



The standard deviation of wins, much like the annual distribution of wins for each team, tends to fluctuate where around every 7 years, the data points hit there lowest value and then grow towards another peak 3-5 years later. Just as before, the most dramatic plummet in values can be attributed to the pandemic during the 2020 season.

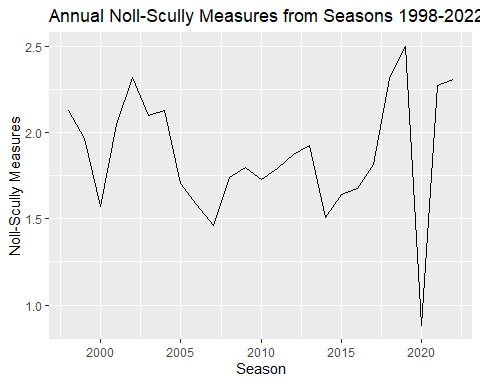
### Task 4:

Calculate the Noll-Scully metric for each season. Plot the Noll-Scully metric to see how it has changed from year to year. Comment on what you see in this plot. Note: Aside from the y-axis, this plot should have an identical appearance to your plot from Task 3.

NSmetric = StddevWins\_graph %>%  
 group\_by(Season = yearID) %>%  
 summarize(NollScully = StdevWins/ISD)  
NSmetric

## # A tibble: 25 × 2  
## Season NollScully  
## <int> <dbl>  
## 1 1998 2.13  
## 2 1999 1.97  
## 3 2000 1.57  
## 4 2001 2.04  
## 5 2002 2.32  
## 6 2003 2.10  
## 7 2004 2.13  
## 8 2005 1.70  
## 9 2006 1.58  
## 10 2007 1.46  
## # ℹ 15 more rows

ggplot(NSmetric,aes(Season,NollScully))+  
 geom\_line() +  
 labs(x="Season",y="Noll-Scully Measures",title="Annual Noll-Scully Measures from Seasons 1998-2022")



As to be expected, this graph mirrors the previous one that measured the annual standard deviation of wins. We know that for a season to have a perfect competitive balance, the Noll-Scully value must be a perfect 1. Similarly, lower NS values are indicative of more competitively balanced years. In our dataset, the 2000,2007, and 2014 seasons had NS values of around 1.5. These were the “best” in the pre-pandemic years that we evaluated, the only NS value to fall below closer to 1 being 2020’s value of 0.88.

### Task 5:

Now your task is to create Noll-Scully plots (versus year) for the National League and American League (separately). You will need to recalculate the standard deviation of wins for each league. Present the two plots in an easy to compare manner and comment on the differences in competitive balance over time.

### Calculations for National League Data:

NL = teams %>%   
 filter(lgID == "NL")

StddevWins\_NL = NL %>%   
 group\_by(yearID) %>%  
 summarize(StdDevWins = sd(W))  
StddevWins\_NL

## # A tibble: 25 × 2  
## yearID StdDevWins  
## <int> <dbl>  
## 1 1998 14.3   
## 2 1999 12.9   
## 3 2000 11.2   
## 4 2001 10.5   
## 5 2002 13.0   
## 6 2003 11.3   
## 7 2004 14.1   
## 8 2005 8.83  
## 9 2006 8.04  
## 10 2007 7.96  
## # ℹ 15 more rows

# Noll-Scolly Measures  
NSmetric\_NL = StddevWins\_NL %>%  
 group\_by(Season=yearID) %>%  
 summarize(NollScully = StdDevWins/ISD)

#esquisser()  
NSMetricNL\_graph = ggplot(NSmetric\_NL) +  
 aes(x = Season, y = NollScully) +  
 geom\_line(colour = "#112446") +  
 labs(  
 title = "National League Annual Noll Scully Values",  
 subtitle = "From 1998-2022"  
 ) +  
 theme\_minimal()

### Calculations for American League Data:

AL = teams %>%  
 filter(lgID == "AL")

StddevWins\_AL = AL %>%   
 group\_by(yearID) %>%  
 summarize(StdDevOfWins = sd(W))  
StddevWins\_AL

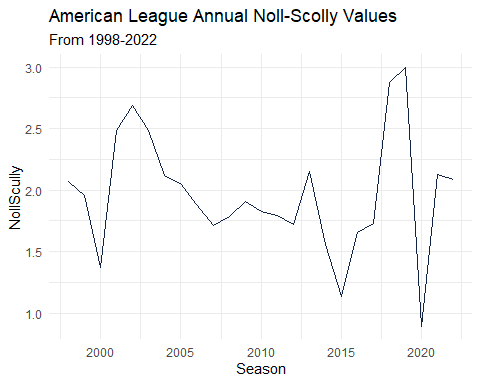
## # A tibble: 25 × 2  
## yearID StdDevOfWins  
## <int> <dbl>  
## 1 1998 13.2   
## 2 1999 12.5   
## 3 2000 8.71  
## 4 2001 15.8   
## 5 2002 17.1   
## 6 2003 15.8   
## 7 2004 13.5   
## 8 2005 13.1   
## 9 2006 12   
## 10 2007 10.9   
## # ℹ 15 more rows

# Noll-Scolly Measures  
NS\_AL = StddevWins\_AL %>%  
 group\_by(Season=yearID)%>%  
 summarize(NollScully= StdDevOfWins/ISD)

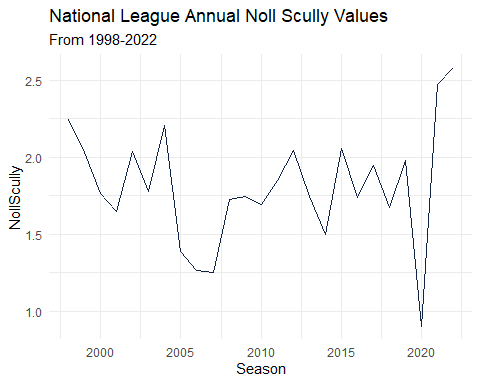
#esquisser()  
NSMetricAL\_graph = ggplot(NS\_AL) +  
 aes(x = Season, y = NollScully) +  
 geom\_line(colour = "#112446") +  
 labs(  
 title = "American League Annual Noll-Scolly Values",  
 subtitle = "From 1998-2022"  
 ) +  
 theme\_minimal()

### Side-by-side Graphs for the Noll-Scurry Values for Both Leagues

par(mfrow = c(1, 2))  
NSMetricAL\_graph

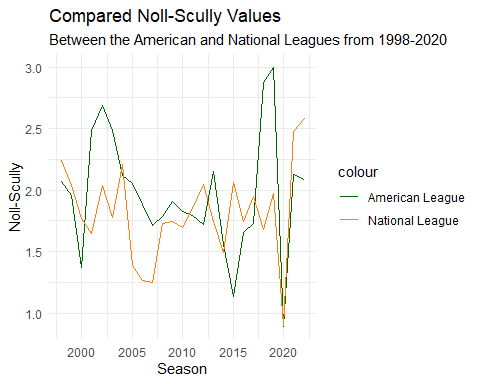


NSMetricNL\_graph



### Overlayed Combined Line Graph for Both Leagues

ggplot() +  
 geom\_line(data = NS\_AL, aes(Season, NollScully, color = "American League")) +  
 geom\_line(data = NSmetric\_NL, aes(Season, NollScully, color = "National League")) +  
 labs(x = "Season", y = "Noll-Scully", title = "Compared Noll-Scully Values", subtitle = "Between the American and National Leagues from 1998-2020") +  
 theme\_minimal() +  
 scale\_color\_manual(values = c("#006400","#ff7f0e"), labels = c("American League", "National League"))



The competitive balance appears to show less overall fluctuation in the the National League teams, with a slight expection being between 2004-2005, and of course the 2020 season. The American League has two peaks that tower over the comparative peaks for the National Teams, but this trend is not consistent throughout the years since there are several years where the Noll-Scully value is higher for the National League teams than the American League Teams.