Homework 3

Narek Sahakyan

27 October, 2019

TOTAL: 50p

Reminder: In football draw is a half win. Number of teams per season might change even within given league

Noll-Skully number

Your goal is to calculate Noll-Skully number for football Chose any 4 leagues from f_data_sm on your own * Calculate Noll_Skully Number as a whole for all the seasons for the given league. (2p)

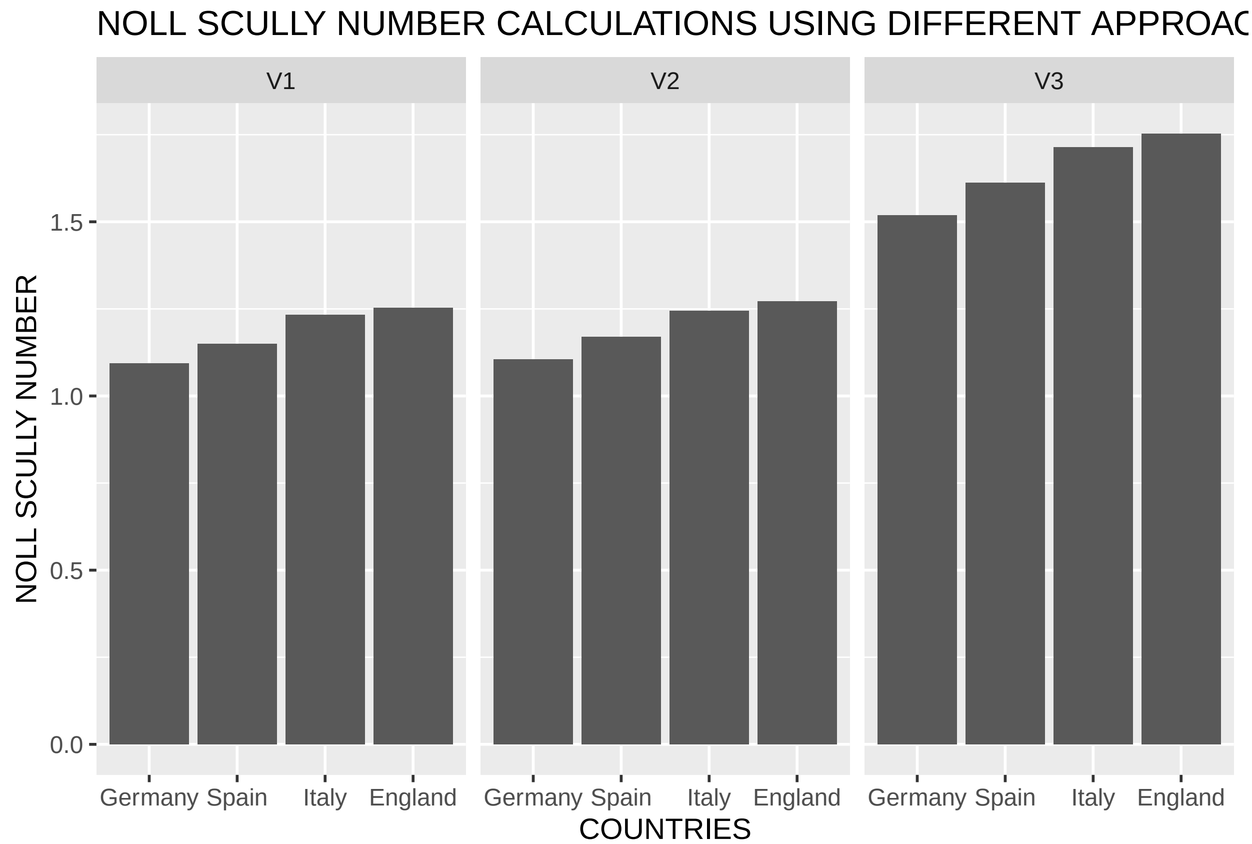
```
countries <- c("England","Germany","Spain","Italy")
top_4_data <- f_data_sm %>%
  filter(COUNTRY %in% countries)

final_tables <- function(data,country){
  result <- c()
  for(season in unique(data$SEASON)){
    season_table <- final_table(data, country, season)
    league <- data[data$SEASON == season &
                     data$COUNTRY == country,]$LEAGUE[1]
    season_table$SEASON = season
    season_table$COUNTRY = country
    season_table$LEAGUE = league
    result <- rbind(result, season_table)
  }
  cols <- colnames(result)
  len <- length(cols)
  result <- result[c("SEASON","COUNTRY","LEAGUE",cols[1 : (len-3)])]
  return(as.data.frame(result))
}

countries_standings <- function(data){
  result <- c()
  for(country in unique(data$COUNTRY)){
    country_standings <- final_tables(data,country)
    result <- rbind(result, country_standings)
  }
  return(result)
}

top_4_standings <- countries_standings(top_4_data)
top_4_standings$TW = round(top_4_standings$W + (top_4_standings$D)/2)
top_4_standings$WR <- top_4_standings$TW / top_4_standings$M
#Win Ratio = number of wins / number of games
#Tunned wins = number of wins + (number of draws)/2
#As already mentioned in the description the number of games played
#each season in football can change even in the same league,
#so in order to have a general metric for summarizing let's take
```

```r
#the mean of the number of games played in each season as an approximation
#We will use the exact number of games later,
#when analyzing the CB using season based approaches

#Version 1 teams_count
#Number of teams = number of games / 2 + 1
# teams_count <- top_4_standings %>%
#   group_by(COUNTRY) %>%
#   summarise(MEAN.TC = round(mean( M/2 + 1 )))
# id_s <- 0.5 / sqrt(teams_count)

mean_t <- function(M) {
  return(round( mean( M/2 + 1 ) ))
}

top_4_nsc_v1 <- top_4_standings %>%
  group_by(COUNTRY) %>%
  summarise(NSC = sd(WR) / ( 0.5 / sqrt(mean_t(M)))) %>%
  mutate(V = "V1") %>%
  arrange(desc(NSC))
#Now let's use count three draws as one win
top_4_standings$TW_v2 <- round(top_4_standings$W + (top_4_standings$D)/3)
top_4_standings$WR_v2 <- top_4_standings$TW_v2 / top_4_standings$M
top_4_nsc_v2 <- top_4_standings %>%
  group_by(COUNTRY) %>%
  summarise(NSC = sd(WR_v2) / ( 0.5 / sqrt(mean_t(M)))) %>%
  mutate(V = "V2") %>%
  arrange(desc(NSC))

# As we can see the competitive balance increased for all the leagues
# when using this type of calculations.
# The increase in the competetive balance
# means that the role of luck gets less decisive.
# However the changes are very slight and
# not significant, so I am not sure,
# but I believe that having these facts we can conclude
# that the draws in general do slightly affect the all time CB for these leagues, as
# when we give them less importance(v1: 2 draws = 1 win, v2: 3 draws = 1 win)
# the CB is increasing which can be translated into
# the decrease of the importance of the luck. I find this connection logical
# as if the CB is getting higher when we discard more draws,
# as we count 3 of them as a win. However when the CB is close to
# being balanced we expect more draws in general.
# Now let's imagine two simple examples.
# Suppose team X won 1 game and tied 10 times.
# If we want to transform the draws into wins using half principle
# We would say that X won 6 of its games, using 1/3 principle
# we would say that the team won 4 of its games.
# According to our approach, in general the discard of that two "won" games leads to
# increase in CB(decrease in luck importance).
# Using 1/3 principle the teams with most wins will get the
# highest portion of wins, and the teams with more draws
# will get lower rates in comparison to the rates
```

```r
# they would get using half principle(6 vs 4) and
# this will lead to higher variance in skill(CB). So the teams
# with draws the lower will become this variance in skill,
# which will increase the luck's importance as in
# general there would be more equal teams and more expected draws.

# I hope I have written something logical :)



#Version 2 matches count
# Let's use the same approach as for version one,
# but considering the number of games played
# rather than the number of the teams.
mean_m <- function(M){
  return(round(mean(M)))
}

top_4_nsc_v3 <- top_4_standings %>%
  group_by(COUNTRY) %>%
  summarise(NSC = sd(WR_v2) / ( 0.5 / sqrt(mean_m(M)))) %>%
  mutate(V = "V3") %>%
  arrange(desc(NSC))

top_4_nsc <- rbind(top_4_nsc_v1,
                   top_4_nsc_v2,
                   top_4_nsc_v3)
# As we can see all the approaches identified the same ranking
# of the leagues in terms of CB over all time, so let's check
# the differences in CB for the leagues for all of the approaches
```

- Plot these numbers together and comment (6p)

```r
top_4_nsc %>%
  ggplot(aes(x = reorder(COUNTRY,NSC))) +
  geom_bar(aes(y = NSC),stat = "identity") +
  facet_grid(.~V) +
  labs(x = "COUNTRIES", y = "NOLL SCULLY NUMBER") +
  ggtitle("NOLL SCULLY NUMBER CALCULATIONS USING DIFFERENT APPROACHES")
```




```r
# As we can see the differences in the leagues in terms of CB balances remain
# almost the same for V1 and V2, but become more sparse for V3.
# However this is mainly because of considering
# the number of matches instead of number of teams.
# I believe that using V3 is more relevant for football.
# As we know in general the
# Number of Games in a competition = (Number of teams - 1 ) * 2,
# So addition of one team will lead to two more games, two teams = 4 more games and so on.
# In general, more the games, more the chances of change in CB.
# Supposing that we calculate ID_S by the formula
# 0.5 / sqrt(FACTOR), where FACTOR
# is either 1)the number of games or 2)the number of teams.
# Unit increase in 1) leads to double of that increase in 2),
# the higher the denominator the lower the ID_S
# which in case will lead to higher CB as CB = SD(WPCT)/ID_S.
# As number of games are derived
# from the number of teams(opposite is also true in general) and in general
# they are more important in CB as CB is more sensitive to it and becomes
# higher when regarding them as factor,
# In my opinion it is better to use number of games as a FACTOR for ID_S

top_4_nsc %>%
  group_by(V) %>%
  summarise(SD = sd(NSC)) %>%
  arrange(desc(SD))
```

```
## # A tibble: 3 x 2
##   V          SD
##   <chr>  <dbl>
## 1 V3    0.106
## 2 V2    0.0757
## 3 V1    0.0740
```

- Now do the same by season (6p)
- Plot and comment (8P)

Your goal is to find 2 leading and 2 lagging indicators for those leagues. Show correlation (on plot and calculating correlation coefficient) between these indicators and Noll-Skully number (or any other competitive balance metric on your choice). 20p

Explain why do you think these variables are leading or lagging. 10p