DA 607: NFL Spreads: Populations, Underdogs, and other Factors

Contents

Introduction

- Vegas bookies create the spreads for NFL Football games.
- They don't care if the spread is "correct", they want 50% of betters to choose each team (guaranteeing their "rake")
- If betters favor certain matchup attributes, perhaps this can be exploited.

Factors to Analyze

- **Historic NFL Lines:** Does a teams historical performance against the spread create bias on the current spread? Example: How often have they covered in the past?
- **Harris Poll Popularity Rankings:** Does a teams "popularity", as measured by the Harris Poll, create betting bias?
- **City Populations:** Does the size of the city, and thus their "betting population", create bias in the spreads?
- **Game Location:** How often does the Home team cover the spread?
- **Points Scored:** Does "how much they covered by" history affect the teams current spreads?
- **Stadium Attendance:** Does the teams stadium attendance (percent of stadium capacity filled) show a betting bias that can be exploited?

ESEMN Workflow

Obtain + Scrub form the following 3 Data Sources

Source 1: Historic NFL Lines

Download the Historic NFL Lines CSVs from web to disk:

https://github.com/mattrjacobs/nflspread/tree/master/files

```
# dont want to keep downloading from web each time:

url_prefix <- 'https://raw.githubusercontent.com/mattrjacobs/nflspread/master/files/nfl'

if(DO_DOWNLOAD_FILES){
  for(i in 1979:2013)
  {
    input_url <- paste(url_prefix,i,"lines.csv", sep="")
    output_file <- paste("lines/nfl",i,"lines.csv", sep="")
    download.file(input_url, destfile=output_file)
  }
}
```

Source 2: Harris Poll - Scrape for Year-over-Year Team Rankings

This includes the "one data transformation operation", removing all equal (=) signs from the rankings table.

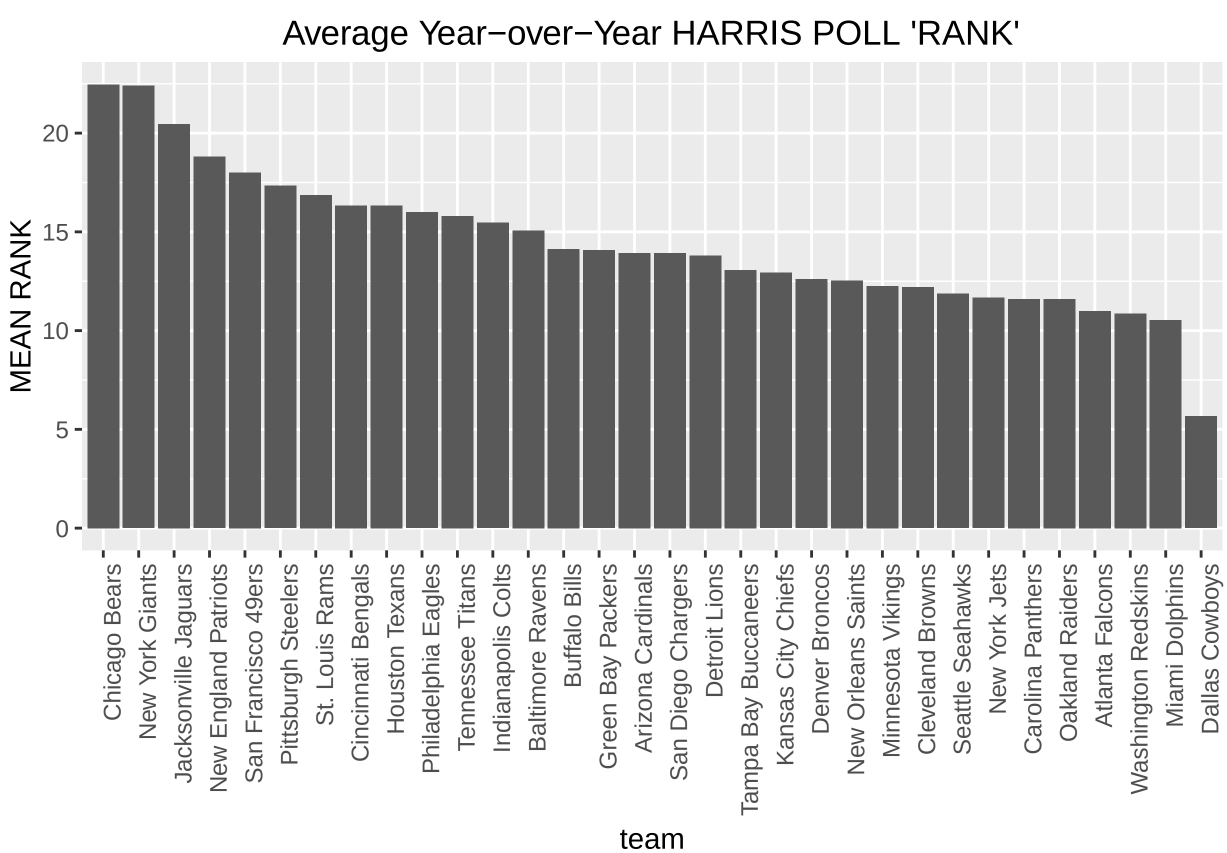
```
if(DO_HARRIS_POLL_FAVS){
 harris_tables <- readHTMLTable(HARRIS_POLL_URL, header = TRUE)
  harris_df <- harris_tables[[3]]
  colnames(harris_df)[1] <- "team"
  # some had = sign before numbers, got it out...
  harris_df <- as.data.frame(sapply(harris_df,gsub,pattern="=",replacement=""))
  numeric_cols <- c(2:ncol(harris_df))
  harris_df[, numeric_cols] <- sapply(harris_df[, numeric_cols], as.numeric)
  harris_df$MEAN_RANK <- rowMeans(subset(harris_df, select = numeric_cols), na.rm = TRUE)
  harris_sub <- c(ncol((harris_df)-5):(ncol(harris_df)-1))
  harris_df$MEAN_RANK_5 <- rowMeans(subset(harris_df, select = harris_sub), na.rm = TRUE)
```

```
  kable(head(harris_df[,-c(4:12)]))
}
```

team	1998	1999	2011	2013	2014	2015	MEAN_RANK	MEAN_RANK_5
Dallas Cowboys	1	1	1	1	22	1	5.666667	3.333333
Green Bay Packers	12	19	20	10	19	11	14.066667	12.533333
New England Patriots	8	3	24	26	26	11	18.800000	14.900000
Denver Broncos	23	22	8	21	1	23	12.600000	17.800000
Pittsburgh Steelers	24	23	10	28	23	24	17.333333	20.666667
Seattle Seahawks	13	7	18	4	24	25	11.866667	18.433333

See above for MEAN of rankings, and below for graphcs of team's mean ranks.

```
the_aes <- aes(x=reorder(team,-MEAN_RANK), y=MEAN_RANK)
plt <- ggplot(harris_df, the_aes) + geom_bar(stat="identity")
plt <- plt + ggtitle("Average Year-over-Year HARRIS POLL 'RANK'") + labs(x="team", y="MEAN RANK")
plt <- plt + stat_summary(fun.y=sum, geom="bar")
plt <- plt + theme(axis.text.x = element_text(angle = 90, hjust = 1))
show(plt)
```



Source 3: City Populations

Data was copy-pasted from the following Wikipedia Link:

https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

and in the following manner:

```
do_explicit_populations <- function(df){

  df[, "pop"] <- 0

  df[df$team == "New York Giants", "pop"] <- 8491079
  df[df$team == "Green Bay Packers", "pop"] <- 104891

  # same for the rest of the teams....not showing here (too many lines...)
}
```

team	pop
Dallas Cowboys	1281047
Green Bay Packers	104891
New England Patriots	655884
Denver Broncos	663862
Pittsburgh Steelers	305412
Seattle Seahawks	668342

Work

In these steps, just gathering the data and raw calculations, will do analysis in "Conclusions")

Work 1: Load the downloaded CSVs of historical games into a Data Frame.

DataFrame created: "lines_df""

```
lines_df <- read.csv(paste("lines/nfl",1978,"lines.csv", sep=""))
lines_df['season'] <- 1978
for(i in 1979:2013)
{
  filepath <- paste("lines/nfl",i,"lines.csv", sep="")
  #print(filepath)
  lines_df_new <- read.csv(filepath)
  lines_df_new['season'] <- i
  lines_df <- rbind(lines_df, lines_df_new)
}

colnames(lines_df) <- c("date","v_team","v_score","h_team","h_score","line","total", "season")
```

Work 2: New Field: "h_covered": Did the HOME team "cover"?

DataFrame updated: "lines_df""

```
lines_df$h_covered <- ((lines_df$v_score + lines_df$line) > lines_df$h_score)
kable(head(lines_df))
```

date	v_team	v_score	h_team	h_score	line	total	season	h_covered
09/01/1979	Detroit Lions	16	Tampa Bay Buccaneers	31	3	30.0	1978	FALSE
09/02/1979	Atlanta Falcons	40	New Orleans Saints	34	5	32.0	1978	TRUE
09/02/1979	Baltimore Colts	0	Kansas City Chiefs	14	1	37.0	1978	FALSE
09/02/1979	Cincinnati Bengals	0	Denver Broncos	10	3	31.5	1978	FALSE
09/02/1979	Cleveland Browns	25	New York Jets	22	2	41.0	1978	TRUE
09/02/1979	Dallas Cowboys	22	St Louis Cardinals	21	-4	37.0	1978	FALSE

```
# Just confirming the rowcounts are good...winners + losers == rowcount
lines_df_home_covered <- sum(lines_df$h_covered==TRUE)
lines_df_home_didnt_cover <- sum(lines_df$h_covered==FALSE)

# confirm it adds up to total number of rows:
sum(lines_df_home_covered, lines_df_home_didnt_cover) == nrow(lines_df)
```

```
## [1] TRUE
```

Work 3: Choose 2013 Season to Analyze

Subset the spreads data, only take 2013 (the most recent):

DataFrame updated: "lines_df""

```
lines_df <- lines_df[lines_df$season == 2013,]
kable(head(lines_df))
```

	date	v_team	v_score	h_team	h_score	line	total	season	h_cover
8175	09/05/2013	Baltimore Ravens	27	Denver Broncos	49	7.5	49.5	2013	FALSE
8176	09/08/2013	New England Patriots	23	Buffalo Bills	21	-10.5	51.5	2013	FALSE
8177	09/08/2013	Tennessee Titans	16	Pittsburgh Steelers	9	6.0	42.0	2013	TRUE
8178	09/08/2013	Atlanta Falcons	17	New Orleans Saints	23	3.5	56.0	2013	FALSE
8179	09/08/2013	Tampa Bay Buccaneers	17	New York Jets	18	-6.0	39.0	2013	FALSE
8180	09/08/2013	Kansas City Chiefs	28	Jacksonville Jaguars	2	-4.5	43.0	2013	TRUE

Work 4: Create 'Winner Points' Data Subset:

DataFrame created: "covered_by_summary_df"

How much did you score when you won?

Its a union of home_points when home_covered and away_points when not(home_covered)

Therefore, your times covered matter (because you only get the points if you covered) and the total points matter (because we're summing up the points on the days that you covered)

```
covered_by_summary_home_away <- sqldf("select h_team as team, sum(h_score) as winner_points from lines_

covered_by_summary_df <- sqldf("select team, sum(winner_points) as winner_points from covered_by_summar
kable(head(covered_by_summary_df))
```

team	winner_points
Chicago Bears	271
Green Bay Packers	212
Denver Broncos	207
Detroit Lions	202
Houston Texans	195
Washington Redskins	186

Work 6: Merge the h_score and v_score logic (turn into times covered)

DataFrame created: "times_covered_df"

Subset the spreads data, only take 2013 (the most recent):

```
covered_count_home_away <- sqldf("select h_team as team, count(*) as times_covered from lines_df where l

times_covered_df <- sqldf("select team, sum(times_covered) as times_covered from covered_count_home_awa

kable(head(times_covered_df[,]))
```

team	times_covered
Houston Texans	12
Chicago Bears	11
Washington Redskins	11
Jacksonville Jaguars	10
Tampa Bay Buccaneers	10
Atlanta Falcons	9

Work 7: Stadium Attendance

```
attendance_df <- read.csv("Stadium_Attendance.csv")
kable(head(attendance_df))
```

team	Stadium	Home.games	Average.attendance	Total.attendance	Capacity.percentag
Green Bay Packers	Lambeau Field	8	78413	627308	107.
Indianapolis Colts	Lucas Oil Stadium	8	66047	528381	104.
San Francisco 49ers	Levi's Stadium	8	70799	566392	103.
Seattle Seahawks	CenturyLink Field	8	69020	552162	103.
Miami Dolphins	Sun Life Stadium	8	67193	537548	102.
Philadelphia Eagles	Lincoln Financial Field	8	69483	555868	102.

Work 8: Populations (and start the merge...)

```
final_df <- do_explicit_populations(covered_by_summary_df)
kable(head(final_df))
```

team	winner_points	pop
Chicago Bears	271	2722389
Green Bay Packers	212	104891
Denver Broncos	207	663862
Detroit Lions	202	680250
Houston Texans	195	2239558
Washington Redskins	186	658893

Work (FINAL MERGE): All into single data frame for Analysis:

```
harris_to_merge <- harris_df[ , which(names(harris_df) %in% c("team","MEAN_RANK","MEAN_RANK_5"))]
final_df <- merge(final_df,times_covered_df,by="team")
final_df <- merge(final_df,harris_to_merge,by="team")
attendance_df_to_merge <- attendance_df[ , which(names(attendance_df) %in% c("team","Capacity.percentag
final_df <- merge(final_df,attendance_df_to_merge,by="team")

kable(head(final_df))
```

team	winner_points	pop	times_covered	MEAN_RANK	MEAN_RANK_5	Capacity.percenta
Arizona Cardinals	133	1537058	3	13.93333	8.466667	98
Atlanta Falcons	179	456002	9	11.00000	7.500000	98
Baltimore Ravens	156	622793	8	15.06667	14.033333	100
Buffalo Bills	141	258703	8	14.13333	12.066667	95
Carolina Panthers	90	809958	6	11.60000	9.800000	100
Chicago Bears	271	2722389	11	22.46667	24.233333	100

The Calculations

```
# final_df$team +
final_lm <- lm(final_df$times_covered ~ final_df$winner_points + final_df$pop + final_df$MEAN_RANK + fi

summary(final_lm)
```

```
##
## Call:
## lm(formula = final_df$times_covered ~ final_df$winner_points +
##     final_df$pop + final_df$MEAN_RANK + final_df$MEAN_RANK_5 +
##     final_df$Capacity.percentage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8586 -0.9031  0.1352  1.0053  4.0644
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.629e+01  7.329e+00   2.223   0.0355 *
```

```
## final_df$winner_points        1.294e-02  8.735e-03   1.482   0.1509
## final_df$pop                  1.605e-07  1.977e-07   0.812   0.4247
## final_df$MEAN_RANK            4.140e-01  1.792e-01   2.311   0.0294 *
## final_df$MEAN_RANK_5         -2.078e-01  1.200e-01  -1.731   0.0957 .
## final_df$Capacity.percentage -1.460e-01  7.569e-02  -1.929   0.0652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.037 on 25 degrees of freedom
## Multiple R-squared:  0.3468, Adjusted R-squared:  0.2162
## F-statistic: 2.655 on 5 and 25 DF,  p-value: 0.04664
```

```
times_covered_ig <- information.gain(times_covered~., final_df)

print(times_covered_ig)
```

```
##                     attr_importance
## team                        1.60742
## winner_points               0.00000
## pop                         0.00000
## MEAN_RANK                   0.00000
## MEAN_RANK_5                 0.00000
## Capacity.percentage         0.00000
```

Conclusions

There are flaws in the logic here:

- The spreads Data is from 2013 while the "favorites"" data 2015
- Picked random number of years for Harris Poll averages (ie 5 and ALL for which years to summarize)
- Not looking at many other factors, such as win streaks, player statistics, etc.
- Some stats are slightly circular: The team ranks may be determined by how often they cover

With those flaws in mind, there seems to be some factors that could be significant:

- There's only a **6% chance** that the stadium capacity % is due to chance.
- **The MEAN RANKS:** The fact that the **OVERALL MEAN RANK** has more significance than the **5 YEAR MEAN RANK** is interesting. Do "winners" go to historically strong teams? If nothing else, it may show that betters care more about "recent history" than "historical tradition".

It would be quite interesting to look at statistics more historically to see if there was a point and time in which these predictions would have been more possible, when all of the indicators were not already in use to create the lines.