Exercise - College Football Rankings and Market Efficiency

This question is based on Fair & Oster (2007) and the related discussion in Chapter 10 of Predicting Presidential Elections and Other Things by Ray Fair. The data used in this exercise are courtesy of Professor Fair. You can download a copy from the data directory of my website as follows:

```
library(tidyverse)
football <- read_csv('https://ditraglia.com/data/fair_football.csv')
football
```

```
# A tibble: 1,582 × 10
   SPREAD     H   MAT   SAG   BIL   COL   MAS   DUN   REC    LV
    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     34     1     7    31    28    17    38    14  0     24
 2     29    -1    34    29    10    41    26    18 33.3   13.5
 3     10    -1   -16   -23   -33     5   -12   -25  8.33 -10.5
 4    -11     1     2    -8    -8    -7    -2    -4  0      3
 5     35    -1    35    35    38    25    25    28 25       5
 6     -2     1    29    36    17    25    20    11 33.3   11.5
 7     11     1    35    39    28    40    30    34 41.7   10
 8     20     1    29    13    12    37    13    26 25       7.5
 9      7     1    40    41    -7    45    36    43 66.7   11.5
10     20    -1    61    37    36    80    51    35 75     11
# i 1,572 more rows
```

Each row of the tibble `football` contains information on a single division I-A American college football game. All of these games were played in 1998, 1999, 2000, or 2001. We have ten weeks of data for each year, beginning in week 6 of the college football season.

Outcome Variable: `SPREAD`

Our goal is to predict `SPREAD`, the *point spread* in a given football game. This variable is constructed as follows. For each game, one of the two teams is *arbitrarily* designated "Team A" and the other "Team B." The point spread is defined as A's final score minus B's final score. For example, in the first row of `football` the value of `SPREAD` is 34. This means that team A scored 34 more points than team B. Again, the designations of A and B are *completely arbitrary*, so `SPREAD` can be positive or negative. The value of `-2` for `SPREAD` in row 6 indicates that the team designated A in that game scored two points *fewer* than team designated B.

Predictor Variables

Home Field Indicator: `H`

The predictor `H` is a categorical variable that equals `1` if team A was the home team, `-1` if team B was the home team, and `0` if neither was the home team as in, e.g. the Rose Bowl.

Computer Ranking Systems: `(MAT, SAG, BIL, COL, MAS, DUN)`

Our next set of predictors is constructed from the following computer ranking systems:

1. Matthews/Scripps Howard (MAT)
2. Jeff Sagarin's *USA Today* (SAG)
3. Richard Billingsley (BIL)
4. *Atlanta Journal-Constitution* Colley Matrix (COL)
5. Kenneth Massey (MAS)
6. Dunkel (DUN)

Fair and Oster (2007) describe these as follows:

> Each week during a college football season, there are many rankings of the Division I-A teams. Some rankings are based on the votes of sports writers, and some are based on computer algorithms ... The algorithms are generally fairly complicated, and there is no easy way to summarize their main differences.

The predictors `MAT, SAG, BIL, COL, MAS` and `DUN` are constructed as the *difference* of rankings for team A minus team B in the week when the corresponding game is scheduled to occur. Suppose, for example, that in a week when Stanford is schedule to play UCLA, Richard Billingsley has Stanford #10 and UCLA #22. The *difference* of ranks is 12. So if Stanford is team A, `BIL` will equal `12` and if Stanford is team B, `BIL` will equal `-12`. To be clear, each of these predictors will be *positive* when the team designated A is *more highly ranked*.

Win-Loss Record: `REC`

Continuing their discussion of computer ranking systems, Fair and Oster (2007) write:

> Each system more or less starts with a team's win-loss record and makes adjustments from there. An interesting system to use as a basis of comparison is one in which only win-loss records are used ... denoted `REC`.

The predictor `REC` is constructed differently from `MAT, SAG, BIL, COL, MAS` and `DUN`. This predictor equals the difference in *percentage of games won* for team A minus team B. For example, returning to the Stanford versus UCLA example, suppose that Stanford has won 80% of its games thus far while UCLA has won 50%. Then `REC` will equal `30` if Stanford is team A and `-30` if Stanford is team B.

Las Vegas Point Spread: `LV`

Our final predictor is `LV`: the Las Vegas line point spread. ESPN defines a point spread as follows:

> Also known as the line or spread, it [a point spread] is a number chosen by Las Vegas and overseas oddsmakers that will encourage an equal number of people to wager on the underdog as on the favorite. If fans believe that Team A is two touchdowns better than Team B, they may bet them as 14-point favorites. In a point spread, the negative value (-14) indicates the favorite and the positive value (+14) indicates the underdog. Betting a -14 favorite means the team must win

> by at least 15 points to cover the point spread. The +14 underdog team can lose by 13 points and still cover the spread.

For example, the value of 24 for `LV` row 1 of `football` indicates that fans believe team A is 24 points better than team B. The fact that a point spread is an *equilibrium value* chosen to balance the quantity of bets for and against a given team has some important economic implications that we will explore below.

Exercises

1. After loading the data from my website, calculate the *home field advantage*. How often does the home team win? How many more points, on average, does the home team score?
2. Run a linear regression *without an intercept* that uses `H` to predict `SPREAD`. Interpret the coefficient estimates, carry out appropriate inference, and summarize the model fit. Why *doesn't* it make sense to include an intercept in this regression, or indeed in *any* regression predicting `SPREAD`?
3. The R package `GGally` has a handy function called `ggpairs()` that can be used to make a very attractive and informative data visualization called a *pairs plot*. Install `GGally` and through a combination of reading the help file for `ggpairs()` and searching the internet, find out how to make such a plot and what information it contains. Then use `ggpairs()` to produce a pairs plot of the columns `MAT, SAG, BIL, COL, MAS, DUN,` and `REC`. Briefly discuss your results.
4. Run a regression *without an intercept* using `H, REC` and the six computer ranking systems (`MAT, SAG, BIL, COL, MAS,` and `DUN`) to predict `SPREAD`. Do all of the ranking systems add additional predictive information beyond that contained in `H` and the other ranking systems? Carry out appropriate statistical inference to make this determination. If, based on your results, some predictors appear to be redundant, re-estimate your regression dropping them. Based on your results, is it possible to make better predictions of college football games than the *best* of the seven computer systems?
5. Run a regression *without an intercept* that predicts `SPREAD` using `LV`, `H` and whichever of the seven ranking systems you found to contain independent information in part 4 above. Does `H` or any of the ranking systems contain additional predictive information beyond that contained in `LV`? Carry out appropriate statistical inference to make this determination.
6. What do your findings from part 5 above have to do with the concept of market efficiency? If betting markets are efficient, what should be the slope in a regression that uses `LV` *alone* to predict `SPREAD`? Can you statistically reject these values for the regression coefficients? How accurately does `LV` alone predict `SPREAD`?
7. Produce a single nicely-formatted table summarizing the results of all the regressions that you ran in this question.