Lab #6 - Predictive Regression II

Econ 224

September 11th, 2018

College Football Rankings and Market Efficiency

This example is based on the paper "College Football Rankings and Market Efficiency" by Ray Fair and John F. Oster (*Journal of Sports Economics*, Vol. 8 No. 1, February 2007, pp. 3-18) 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. For convenience I have posted a copy on the course website which can be read into R as follows:

```
library(tidyverse)
football <- read_csv('http://ditraglia.com/econ224/fair_football.csv')
football
```

```
# A tibble: 1,582 x 10
   SPREAD     H   MAT   SAG   BIL   COL   MAS   DUN   REC    LV
    <int> <int> <int> <int> <int> <int> <int> <int> <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  
# ... with 1,572 more rows
```

Each row of the tibble `football` contains information on a single division I-A 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.

Response 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 11. So if Stanford is team A, `BIL` will equal `11` and if Stanford is team B, `BIL` will equal `-11`. 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. 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. Install the R package `GGally` and use the function `ggpairs` to make a pairs plot of the columns `MAT, SAG, BIL, COL, MAS, DUN,` and `REC`. Summarize your results, including the numeric values included in the plot.
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 these. Based on your results from part 4 of this question, 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`?

Solutions

Exercise #1

The home team wins approximately 58.6% of the time which is an 8.6% advantage. On average, the home team scores about 4.86 more points than the away team.

```r
football %>%
  filter(H != 0) %>%
  summarize(Hwin = mean(SPREAD * H > 0), Hpoints = mean(SPREAD * H))
```

```
# A tibble: 1 x 2
   Hwin Hpoints
  <dbl>   <dbl>
1 0.586    4.86
```

Exercise #2

It doesn't make sense to include an intercept since the choice of which team was designated team A was *arbitrary*. The regression intercept is the prediction we should make if all of the predictors were zero. In our example, having all the predictors equal to zero means that neither team is expected to have an advantage over the other so our model should predict a `SPREAD` of zero in this case. Excluding an intercept ensures that this is precisely what it does.

Since we have excluded the intercept, the coefficient on `H` is *precisely* the average number of additional points that the home team scores, relative to the away team. This matches our calculations from Exercise #1 above, but using the regression output we can also carry out inference. The regression provides overwhelmingly strong statistical evidence of a home field advantage.

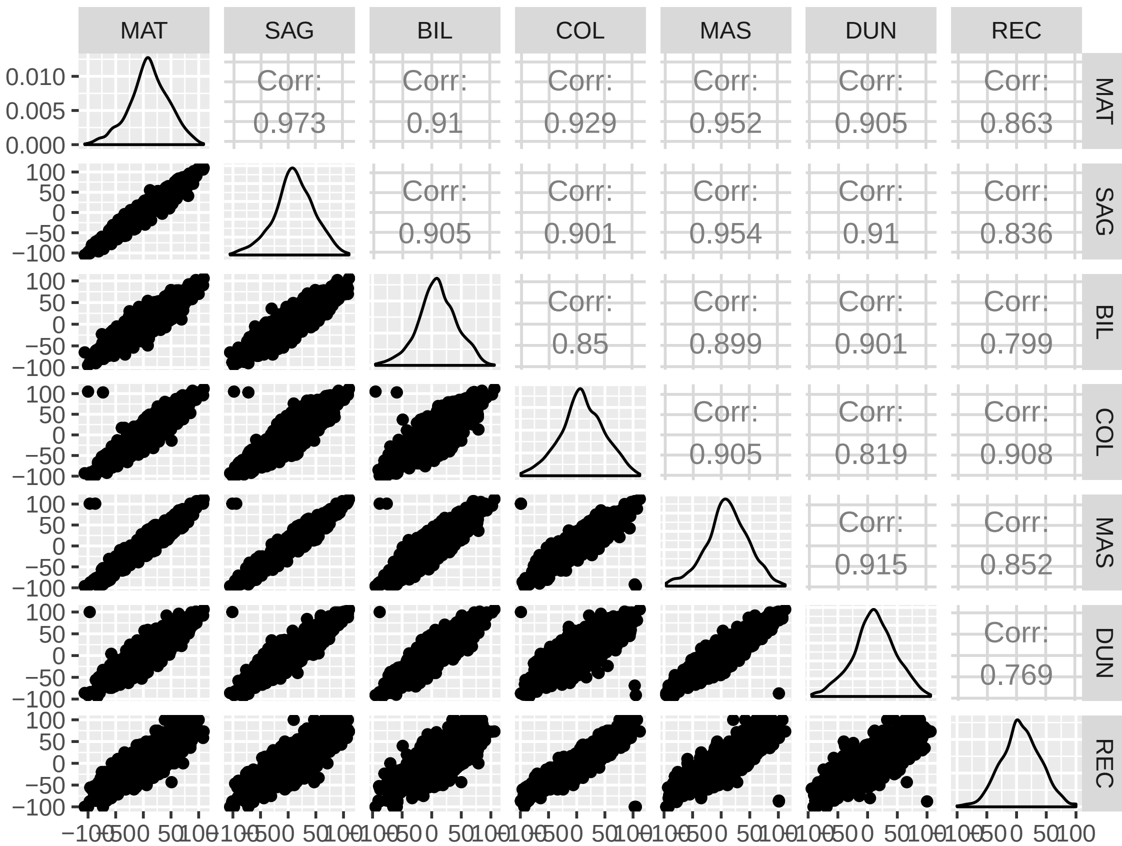
```
reg1 <- lm(SPREAD ~ H - 1, football)
summary(reg1)
```

```
Call:
lm(formula = SPREAD ~ H - 1, data = football)

Residuals:
    Min      1Q  Median      3Q     Max
-61.143  -6.143   6.143  17.857  68.143

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
H    4.857      0.537   9.044   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.66 on 1581 degrees of freedom
Multiple R-squared:  0.04919,   Adjusted R-squared:  0.04859
F-statistic:  81.8 on 1 and 1581 DF,  p-value: < 2.2e-16
```

Exercise #3

These predictors are highly positively correlated, which makes sense: all of them are constructed by starting with `REC` and making adjustments from there. However, the correlations are far from perfect. For example, the correlation between `REC` and `DUN` is only 0.769.

```
library(GGally)
football %>%
  select(MAT:REC) %>%
  ggpairs
```



Exercise #4

Neither `MAT` nor `MAS` are statistically significant taken individually. Moreover, we cannot reject the null hypothesis that these two variables are jointly irrelevant for predicting `SPREAD`:

```
reg2 <- lm(SPREAD ~ H + MAT + SAG + BIL + COL + MAS + DUN + REC - 1, football)

summary(reg2)
```

```
Call:
lm(formula = SPREAD ~ H + MAT + SAG + BIL + COL + MAS + DUN +
    REC - 1, data = football)

Residuals:
    Min      1Q  Median      3Q     Max
-53.542  -9.134   2.150  11.736  56.963

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
H    4.267073   0.436668   9.772  < 2e-16 ***
MAT -0.099306   0.060804  -1.633 0.102624
SAG  0.248165   0.054817   4.527 6.43e-06 ***
BIL  0.080436   0.034244   2.349 0.018953 *
COL -0.062588   0.035894  -1.744 0.081410 .
```

```
MAS -0.007075   0.044624  -0.159 0.874047
DUN  0.118512   0.033769   3.509 0.000462 ***
REC  0.080412   0.030460   2.640 0.008374 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.53 on 1574 degrees of freedom
Multiple R-squared:  0.3942,    Adjusted R-squared:  0.3911
F-statistic:   128 on 8 and 1574 DF,  p-value: < 2.2e-16
```

```r
library(car)
linearHypothesis(reg2, c('MAT = 0', 'MAS = 0'))
```

```
Linear hypothesis test

Hypothesis:
MAT = 0
MAS = 0

Model 1: restricted model
Model 2: SPREAD ~ H + MAT + SAG + BIL + COL + MAS + DUN + REC - 1

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1576 430638
2   1574 429879  2    758.07 1.3878 0.2499
```

This suggests that, after controlling for the other predictors, `MAT` and `MAS` do not add any additional predictive information. Estimating a model without them gives the following results:

```r
reg3 <- lm(SPREAD ~ H + SAG + BIL + COL + DUN + REC - 1, football)
summary(reg3)
```

```
Call:
lm(formula = SPREAD ~ H + SAG + BIL + COL + DUN + REC - 1, data = football)

Residuals:
    Min      1Q  Median      3Q     Max
-53.379  -9.159   2.226  11.953  60.007

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
H    4.31812    0.43495   9.928  < 2e-16 ***
SAG  0.18662    0.03809   4.899 1.06e-06 ***
BIL  0.07203    0.03387   2.127 0.033587 *
COL -0.08575    0.03279  -2.615 0.009014 **
DUN  0.10866    0.03151   3.449 0.000578 ***
REC  0.07666    0.03017   2.541 0.011141 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.53 on 1576 degrees of freedom
Multiple R-squared:  0.3932,    Adjusted R-squared:  0.3908
F-statistic: 170.2 on 6 and 1576 DF,  p-value: < 2.2e-16
```

This model predicts to an accuracy of approximately 16.53 points. This is slightly better than the best individual model, which uses only `H` and `SAG`.

```
reg4 <- lm(SPREAD ~ H + SAG - 1, football)
summary(reg4)
```

```
Call:
lm(formula = SPREAD ~ H + SAG - 1, data = football)

Residuals:
    Min      1Q  Median      3Q     Max
-54.991  -9.074   1.933  12.306  58.642

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
H    4.13374    0.43500   9.503   <2e-16 ***
SAG  0.31952    0.01104  28.936   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.71 on 1580 degrees of freedom
Multiple R-squared:  0.3785,    Adjusted R-squared:  0.3777
F-statistic: 481.2 on 2 and 1580 DF,  p-value: < 2.2e-16
```

However, we strongly reject the null hypothesis that `BIL`, `COL`, `DUN` and `REC` are redundant after including `H` and `SAG`:

```
linearHypothesis(reg3, c('BIL = 0', 'COL = 0', 'DUN = 0', 'REC = 0'))
```

```
Linear hypothesis test

Hypothesis:
BIL = 0
COL = 0
DUN = 0
REC = 0

Model 1: restricted model
Model 2: SPREAD ~ H + SAG + BIL + COL + DUN + REC - 1

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1580 441021
2   1576 430638  4     10384 9.5003 1.372e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise #5

After controlling for `LV` all of the other predictors are irrelevant:

```r
reg5 <- lm(SPREAD ~ LV + H + SAG + BIL + COL + DUN + REC - 1, football)
summary(reg5)
```

```

Call:
lm(formula = SPREAD ~ LV + H + SAG + BIL + COL + DUN + REC - 
    1, data = football)

Residuals:
    Min      1Q  Median      3Q     Max 
-60.379  -8.469   1.564  11.285  54.636 

Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
LV   1.051782   0.076071  13.826   <2e-16 ***
H    0.729503   0.485981   1.501    0.134    
SAG  0.018065   0.037994   0.475    0.635    
BIL -0.027867   0.032797  -0.850    0.396    
COL -0.005476   0.031518  -0.174    0.862    
DUN -0.024891   0.031290  -0.795    0.426    
REC  0.018585   0.028804   0.645    0.519    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.61 on 1575 degrees of freedom
Multiple R-squared:  0.4588,    Adjusted R-squared:  0.4564 
F-statistic: 190.8 on 7 and 1575 DF,  p-value: < 2.2e-16
```

```r
linearHypothesis(reg5, c('H = 0', 'SAG = 0', 'BIL = 0', 'COL = 0', 'DUN = 0', 'REC = 0'))
```

```

Linear hypothesis test

Hypothesis:
H = 0
SAG = 0
BIL = 0
COL = 0
DUN = 0
REC = 0

Model 1: restricted model
Model 2: SPREAD ~ LV + H + SAG + BIL + COL + DUN + REC - 1

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1581 385883                           
2   1575 384026  6    1856.8 1.2692 0.2684
```

Exercise #6

If betting markets are efficient, then `LV` should already contain all available information that would be helpful for predicting `SPREAD`. In line with the theory of market efficiency, we found all of the other predictors to be *redundant* in Exercise #5 above. If betting markets are efficient, `LV` should also provide an *unbiased* prediction of SPREAD so that in the regression SPREAD $= \beta$LV $+ \epsilon$, β should equal 1. If this were not the case, we could use historical data and linear regression to work out the true coefficient values and use this information to win money. But precisely because we have such an incentive to bet when if LV gets out of line with available information, any such anomalies should disappear quickly. Indeed, the estimate of β is very close to 1 and we cannot reject the null hypothesis that it is equal to 1. The Las Vegas line predicts to an accuracy of about 15.6 points.

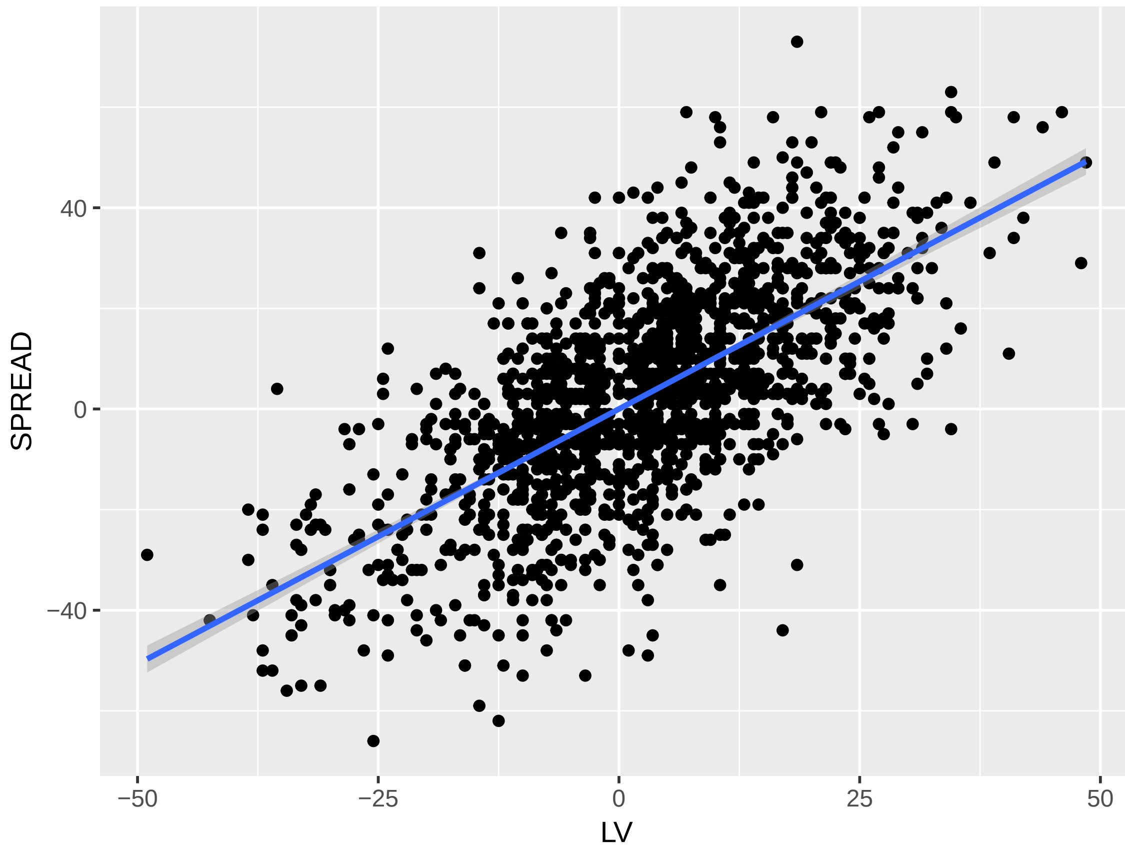
```r
reg6 <- lm(SPREAD ~ LV - 1, football)
summary(reg6)
```

```

Call:
lm(formula = SPREAD ~ LV - 1, data = football)

Residuals:
    Min      1Q  Median      3Q     Max 
-61.244  -9.065   1.043  10.910  54.234 

Coefficients:
   Estimate Std. Error t value Pr(>|t|)    
LV  1.01436    0.02785   36.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.62 on 1581 degrees of freedom
Multiple R-squared:  0.4562,	Adjusted R-squared:  0.4559 
F-statistic:  1326 on 1 and 1581 DF,  p-value: < 2.2e-16
```

```r
linearHypothesis(reg6, c('LV = 1'))
```

```

Linear hypothesis test

Hypothesis:
LV = 1

Model 1: restricted model
Model 2: SPREAD ~ LV - 1

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1582 385948                           
2   1581 385883  1    64.908 0.2659 0.6061
```

```r
ggplot(football, aes(x = LV, y = SPREAD)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x - 1)
```



SPREAD
40
0
−40
−50
−25
0
25
50
LV