If the Vegas line is always moving, what factors are held fixed?

Mary Washington

ECON 360

Dr. Mark Witte

01 December 2018

ABSTRACT

In this paper, the aspects that do not move the Vegas line, are examined using linear regression. 232 observations of college footballs games, from 2004 till 2007, are used to determine what Vegas chooses to disregard in their decision when making a dynamic line. The author has chosen to use a previously mined dataset to investigate what might be overlooked in the construction of the Vegas line. Three, uncorrelated factors, are the apparent stand of the inconsequential factors, not used in making this line move.

1 INTRODUCTION

According to Merriam-Webster, dynamic means to be marked by usually continuous and productive activity or change. The Vegas line is dynamic and adjusts according to wagers placed and updated player and team information. It is constantly impacted by changes that happen hourly in the world of sports. The emphasis that Vegas places on certain factors, is then reflected by the fluidity of the line. According to previous literature, by Witte (2016) and Mirabile (2016)[1], Vegas uses a multi-factor approach in deciding on how to determine the true relationship between the margin of victory and the actual Vegas line. I will be using the estimation of that multi-factored approach and then negating those results, in order to determine those factor that have gotten overlooked. The object of this study was to forecast how Vegas determines their line and then whittle it down to which factors were then ignored.

Using various statistical indicators, from 2004 until 2007, the line for NCAA Football Bowl (FBS) will be determined. Variable selection processes will be used to find the set of regressors that make the best prediction in the line's shift, as well as the complement of that set of regressors that will mark what is discounted in the movement of the Vegas line. 29 indicators are used while one response variable is used, the Vegas line. Ordinarily, using a 5% significance level is appropriate to determine being a valid model, but since the opposite is needed, a failure to reject the null hypothesis is needed. Three regressors were determined to not be statistically

1. Witte, Mark David, and Mirabile, McDonald Paul, 2016. College Football and the Vegas Line: Deconstruction and Arbitrage, *The Journal of Prediction Markets*, Vol. 10, No. 1, 53-67

significant in the adjustment of the Vegas line. First, not being a conference member from either SEC, ACC, Big Ten, Big 12, or Pac-12, matters for line stability. Second, the teams must not be a rival of each other. Lastly, the favored team must possess no injuries. This suggests that the line does not move when all of these factors are in play, and it will be the same regardless of when one decides to bet on that game.

2 DATA AND METHODOLOGY

The dataset that was used was compiled by Dr. Mark David Witte and Dr. McDonald Paul Mirabile[1], and used with permission. The home and away rating from Jeff Sagarin was also used in this dataset. He has devised a way to account for the team's schedule difficulty and location and used that as a ranking, and then take in the individual skills of each player and classify those as predictors (which makes unbiased by using the Elo rating system). The Elo rating system was developed by Arpad Elo and he took the difference in ratings between two players, then that decided the outcome of the game. Points are then transferred from player to player in accordance to the game's outcome. This system is self-correcting, and thus unbiased. The predictors and rankings are combined to form a rating system based on the teams either playing at home or away.

Many descriptive statistics were used in forecasting the Vegas line. There were 29 predictors in the full model. Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) was then used to do a backwards selection of the best forecast of the Vegas line. This model selection was used multiple times in conjunction with the variance inflation factor to ensure there was no multicollinearity. Both criterions picked the same end model but best adjusted R^2 was used to decided between them. The predictors that were left out of the initial selection process was also added in since those were the ones that did a poor job of predicting the movement of the line, which was what was needed.

Table: Descriptive Statistics

Variable Name	Variable description	mean	Std Dev
line	Vegas line	9.634199134	7.476147528
etp	Estimated total points	51.20779221	8.652475387
fro1	Favored rushing offense	167.7665694	56.11390261
fpo1	Favored passing offense	228.0517811	59.73622551
fto1	Favored total offense	396.044032	63.40272768
fso1	Favored scoring offense	29.53726519	7.566943138
frd1	Favored rushing offense	133.8869726	42.74807975
fped1	Favored pass efficiency defense	119.3015477	18.67959866
ftd1	Favored total defense	347.1926294	67.89269907

fsd1	Favored scoring defense	22.49543598	7.482748225
fsched	Favored opponent's winning percentage	0.532153634	0.118489412
uro1	Underdog rushing offense	144.0243921	51.46881055
upo1	Underdog passing offense	211.0582049	54.86656342
uto1	Underdog total offense	354.7805674	61.41161399
uso1	Underdog scoring offense	24.23914443	6.953074058
urd1	Underdog rushing defense	154.9033671	43.65709919
uped1	Underdog pass efficiency defense	127.4525648	18.06079599
utd1	Underdog total defense	373.9055699	64.79572458
usd1	Underdog scoring defense	26.24521175	7.945179218
usced	Underdog opponent's winning percentage	0.510413507	0.11724282
frank	Favored rank	7.831168831	11.12623527
urank	Underdog rank	3.961038961	9.449787058
fnr	Favored not ranked	0.532467532	0.500028232
unr	Underdog not ranked	0.813852814	0.390070623
finj	Favored injuries	4.489177489	2.812799119
uinj	Underdog injuries	4.372294372	2.895260508
power5	Member of power 5 conferences	0.212121212	0.409697983
AnyRivalry	Any rivalries between two teams	0.155844156	0.363495173
Home_Rating	Jeff Sagarin home rating	71.83099567	11.94332575
Away_Rating	Jeff Sagarin away rating	71.93627706	11.77119391

Full model :

Line=fro1+fpo1+fto1+tso1+frd1+fped1+ftd1+fsd1+fsched+uro1+upo1+uto1+uso1+urd1+uped1+utd1+usd1+usced+power5+AnyRivalry+frank+urank+fnr+unr+Home_Rating+Away_Rating+etp+finj+uinj

First the best prediction for the Vegas line needed to be forecasted, so the full model was used. Backwards selection using AIC and BIC was used to find the best variables for model selection. A summary of then used between the two to see which one had a better adjusted R^2.

Below, are the results from using AIC backwards selection:

```
Step:  AIC=697.28
line ~ fso1 + frd1 + fsd1 + fsched + uto1 + uso1 + urd1 + usd1 +
   usced + frank + urank + fnr + unr + Home_Rating + Away_Rating +
   etp

          Df Sum of Sq    RSS    AIC
<none>                  4079.6 697.28
- urd1       1     63.64 4143.3 698.86
- fsched     1     74.98 4154.6 699.49
- fsd1       1     94.31 4173.9 700.56
- etp        1    122.20 4201.8 702.10
- frank      1    122.57 4202.2 702.12
- fso1       1    135.98 4215.6 702.85
- uto1       1    142.02 4221.6 703.18
- urank      1    161.07 4240.7 704.22
- usd1       1    170.82 4250.4 704.75
- frd1       1    188.30 4267.9 705.70
- uso1       1    207.03 4286.6 706.71
- unr        1    279.22 4358.8 710.57
- usced      1    285.22 4364.8 710.89
- Home_Rating 1    383.03 4462.6 716.01
- fnr        1    423.75 4503.4 718.11
- Away_Rating 1    481.60 4561.2 721.06
```

Below, is the result from using BIC backwards selection:

```
Step:  AIC=749.96
line ~ fso1 + frd1 + uto1 + uso1 + usd1 + usced + frank + urank +
   fnr + unr + Home_Rating + Away_Rating

          Df Sum of Sq    RSS    AIC
<none>                  4370.9 749.96
- uto1       1    109.05 4479.9 750.21
- frank      1    180.65 4551.5 753.87
- uso1       1    184.32 4555.2 754.06
- urank      1    187.25 4558.1 754.21
- usced      1    294.05 4664.9 759.56
- frd1       1    306.72 4677.6 760.19
- unr        1    337.59 4708.5 761.71
- Away_Rating 1    518.24 4889.1 770.40
- fnr        1    603.00 4973.9 774.37
- Home_Rating 1    625.90 4996.8 775.43
- fso1       1    673.43 5044.3 777.62
- usd1       1    880.14 5251.0 786.90
```

3 RESULTS AND DISCUSSION

After deciding the best models from the two Information Criterions, their descriptive statistics were then examined.

Results from backwards selection using AIC:

```
lm(formula = line ~ urd1 + fsched + fsd1 + etp + frank + fso1 +
    uto1 + urank + usd1 + frd1 + uso1 + unr + usced + Home_Rating +
    fnr + Away_Rating, data = ncaaf)

Residuals:
    Min      1Q  Median      3Q     Max
-10.8464 -2.5981 -0.1613  2.8572 14.2336

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.074231   5.970049   1.855 0.064977 .
urd1         0.018328   0.010031   1.827 0.069083 .
fsched       6.191479   3.121999   1.983 0.048626 *
fsd1        -0.175779   0.079031  -2.224 0.027181 *
etp          0.199526   0.078807   2.532 0.012065 *
frank       -0.118088   0.046572  -2.536 0.011939 *
fso1         0.190763   0.071426   2.671 0.008149 **
uto1        -0.025673   0.009406  -2.729 0.006873 **
urank        0.199889   0.068769   2.907 0.004037 **
usd1         0.214073   0.071515   2.993 0.003083 **
frd1        -0.033057   0.010518  -3.143 0.001910 **
uso1        -0.308263   0.093542  -3.295 0.001150 **
unr          6.930608   1.810925   3.827 0.000170 ***
usced      -10.863795   2.808620  -3.868 0.000146 ***
Home_Rating  0.195647   0.043648   4.482 1.20e-05 ***
fnr         -6.548678   1.388999  -4.715 4.36e-06 ***
Away_Rating -0.203135   0.040415  -5.026 1.06e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.366 on 214 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.6827,      Adjusted R-squared:  0.6589
F-statistic: 28.77 on 16 and 214 DF,  p-value: < 2.2e-16
```

The only variable that is statistically insignificant is fsched and the adjusted R^2 for this model is 65.89%. I then compared this to the summary from the model selection using BIC, which is below:

```
lm(formula = line ~ uto1 + frank + uso1 + urank + usced + frd1 +
    unr + Away_Rating + fnr + Home_Rating + fso1 + usd1, data = ncaaf)

Residuals:
    Min      1Q  Median      3Q     Max
-12.5622 -2.7814 -0.2685  2.6450 14.9647

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 13.049444   5.840453   2.234 0.026477 *
uto1        -0.022126   0.009488  -2.332 0.020607 *
frank       -0.139645   0.046523  -3.002 0.002998 **
uso1        -0.272290   0.089806  -3.032 0.002724 **
urank        0.214003   0.070028   3.056 0.002523 **
usced      -11.021457   2.877975  -3.830 0.000168 ***
frd1        -0.034460   0.008811  -3.911 0.000123 ***
unr          7.556512   1.841545   4.103 5.75e-05 ***
Away_Rating -0.193766   0.038113  -5.084 7.95e-07 ***
fnr         -7.438328   1.356359  -5.484 1.15e-07 ***
Home_Rating  0.223351   0.039975   5.587 6.84e-08 ***
fso1         0.266841   0.046043   5.795 2.37e-08 ***
usd1         0.356038   0.053738   6.625 2.67e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 4.478 on 218 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared: 0.66,   **Adjusted R-squared: 0.6413**
F-statistic: 35.26 on 12 and 218 DF, p-value: < 2.2e-16

Between the two the AIC model produced a better adjusted R^2 so that was the model that was chosen. Then all of the significant predictors were discarded and the ones that didn't make the original cut for the model were then added back in. Summaries were then running multiple times, with each predictor that was useful being thrown out, until no predictors were useful. The subsequent steps are below:

lm(formula = **line ~ fped1 + utd1 + power5 + AnyRivalry + finj +
   uinj**, data = ncaaf)

Residuals:
```
   Min      1Q Median     3Q    Max
-12.709  -4.933  -1.372  3.535 27.034
```

Coefficients:
```
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.12969   4.24944   1.678 0.094782 .
fped1      -0.08744   0.02672  -3.272 0.001235 **
utd1        0.02960   0.00784   3.776 0.000205 ***
power5      0.85598   1.30931   0.654 0.513931
AnyRivalry  0.59686   1.30348   0.458 0.647469
finj       -0.03552   0.18154  -0.196 0.845034
uinj        0.40087   0.16414   2.442 0.015367 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 7.159 on 224 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared: 0.1069,   **Adjusted R-squared: 0.08293**
F-statistic: 4.467 on 6 and 224 DF, p-value: 0.0002715

Again,

```
lm(formula = line ~ power5 + AnyRivalry + finj, data = ncaaf)

Residuals:
   Min    1Q Median    3Q    Max
-9.379 -6.071 -2.440  3.968 26.121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.74665    0.96329  10.118   <2e-16 ***
power5       0.70057    1.30488   0.537    0.592
AnyRivalry   0.08917    1.36477   0.065    0.948
finj        -0.06125    0.19010  -0.322    0.748
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.52 on 227 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.001362,   Adjusted R-squared:  -0.01184
F-statistic: 0.1032 on 3 and 227 DF,  p-value: 0.9581
```

This was the final model of the predictors that do adjust the Vegas line.

Line = power5 + AnyRivalry + finj

Lastly, the variance inflation factor was calculated for the final model.

```
power5 AnyRivalry     finj
 1.162334   1.000867  1.162852
```

There is no multicollinearity as the numbers are all pretty low, and well below the acceptable cutoff of 5.

Since power5, any rivalry, and favored team injuries, were the most significant in not affecting the movement of the Vegas line, their mean and standard deviation were examined. The mean of all teams in the dataset having any rivalry was 0.156 and the standard deviation was 0.363. This suggests that most teams of the dataset, doesn't have a rivalry against each other. The favored team having an injury has a mean of 4.489, with a standard deviation of 2.813. This suggests that the teams have an average of 5 people being injured during the season, as half a person cannot be injured. Being a power 5 conference member has a mean of 0.212 and a standard deviation of 0.41. This shows that less than a quarter of the teams are not a member of the top 5 conferences.

4 CONCLUSIONS

The purpose of this investigation was to determine what Vegas does factor in, when they are determining their line, for each college football game. Since the lines are not static and they shift according to new information, what doesn't cause it to adjust was important. My results show that though Vegas does take into account a lot of different criteria for its line (according to Witte and Mirabile) there are aspects that they don't give much thought. These are those teams that are not a member of the top 5 dominant (currently) conferences (SEC, ACC, PAC-12, Big Ten, and Big-12). The team that is favored to win, doesn't have any injuries. Neither the team that is favored, nor the underdog, are rivals of each other. If one looks at a game with these 3 factors, as soon as the line is made, it has a better chance of being more static, when it is again looked at on the day of the game. Essentially these are the games that should be bet upon, and one can wait until the last minute to bet, and the line will be similar to when it was first created.

As for future work, a more current and thus relevant dataset would be ideal. More observations are needed, since 232 is not a lot. A more in depth look at why these factors are not taken into account, is also needed. More years should also be incorporated. Then a comparison between the teams that play against the same teams, each year, can be formed.