Homework 1 – College Football Line and Outcomes Database

Data Reading and manipulation

FIRST, I DROP ALL THE -999 OBSERVATIONS.

A1. What percentage of games is won by the underdog**? A. IF(FMINUSU<0,1,0) THEN TAKE THE AVERAGE OF THAT. 32.6%** What percentage of games is won by the underdog when the favored team is favored by greater than or equal to 10 points? **A. IF(FMINUSU<0,IF(LINE>=10,1,0),0) 9.9%** What percentage of games is won by the underdog when the favored team is favored by less than 10 points? **A. IF(FMINUSU<0,IF(LINE<10,1,0),0) 22.7%**

A2. What percentage of favorites have greater rushing offense than the underdog? **A. IF(fro1-uro1>0,1,0) 66.3%** What percentage of favorites have greater passing offense than the underdog? **A. IF(fpo1-upo1>0,1,0) 61.5%** Which of the on-the-field performance differential between favorite and underdog best predicts which team will be the favorite: rushing offense, passing offense, total offense, scoring offense, rush defense, pass efficiency defense, total defense or scoring defense? **A. MANY WAYS TO DO THIS. I TOOK THE SAME IF STATEMENT FROM ABOVE AND USED IT MULTIPLE TIMES. REMEMBER THAT LOWER NUMBERS ARE BETTER FOR DEFENSE STATS. RUSHING O = 66.3%, PASSING O = 61.5%, TOTAL O = 70.6%, SCORING O = 72.7%, RUSHING D = 65.7%, PASS EFFICIENCY = 65.3%, TOTAL D = 68.1%, SCORING D = 71.0%.**

A3. Create a new variable, labeled "Fans_4_Fav", that it is equal to the stadium capacity when the home team is favored but is negative when the away team is favored. For example, the Penn State-Iowa game (order2=2281) the variable should be 107,282 while the Eastern Michigan-Navy game (order2=1824) the variable should be -30,200. **A. IF("home"="favored",stadium,-1*stadium)**

A4. Create a new dummy variable, labeled "Beaten_up", that is equal to one if the favorite team has more injuries than the underdog and is otherwise equal to zero. What is the average of this variable? (Hint: Are you sure you are only including games where this information is available?) **A. IF(finj>uinj,1,0) 47.6%**

Regression Analytics

B1. What's the R-squared of a simple regression with the score differential outcome as the dependent variable (Y) and the Vegas line as the independent variable (X)? What does the R-squared statistic mean here? Is the Vegas line statistically significant?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.436212							
R Square	0.190281							
Adjusted R	0.186745							
Standard E	14.9017							
Observatic	231							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	1	11950.02	11950.02	53.8142	3.79E-12			
Residual	229	50851.91	222.0607					
Total	230	62801.93						
	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.54418	1.601443	-0.96424	0.335941	-4.69963	1.611266	-4.69963	1.611266
line	0.964145	0.13143	7.335817	3.79E-12	0.705179	1.223112	0.705179	1.223112

R-squared says that 19% of the variation in fminusu can be predicted by the vegas line. T-stat on the vegas line is very large and the coefficient is very significant

B2. What's the R-squared of a simple regression with the score differential outcome as the dependent variable (Y) and the "Fans_4_Fav" as the independent variable (X)? How does the R-squared stat compare to the Vegas line regression and why? Is "Fans_4_Fav" statistically significant?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.093767							
R Square	0.008792							
Adjusted R	0.004464							
Standard E	16.48736							
Observatic	231							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	1	552.1672	552.1672	2.031273	0.155453			
Residual	229	62249.76	271.833					
Total	230	62801.93						
	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	7.245987	1.139805	6.357219	1.1E-09	5.000142	9.491832	5.000142	9.491832
fansforfav	2.62E-05	1.84E-05	1.425227	0.155453	-1E-05	6.25E-05	-1E-05	6.25E-05

Whoa! R-squared drops by a ton! The vegas line is a much better predictor of the outcome of the game relative to fan support for the favorite. T-stat is low on fans for fav. The coefficient is only significant at 1-pvalue = 84.5% confidence interval. Blech.

B3. Using the best on-the-field performance measure that predicted the favorite (from question A2), run another regression with the score differential outcome as the dependent variable (Y) and the best on-the-field performance as the independent variable (X). How does the R-squared stat compare to the Vegas line regression and why? Is the on-the-field performance measure statistically significant?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.101872							
R Square	0.010378							
Adjusted R Sq	0.006056							
Standard Erro	16.47417							
Observations	231							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	1	651.7537	651.7537	2.401467	0.122602			
Residual	229	62150.18	271.3982					
Total	230	62801.93						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	5.03125	2.059271	2.443219	0.015314	0.97371	9.08879	0.97371	9.08879
favscor_o	3.753181	2.421928	1.549667	0.122602	-1.01893	8.525293	-1.01893	8.525293

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.145906							
R Square	0.021289							
Adjusted R Squa	0.017015							
Standard Error	16.3831							
Observations	231							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	1	1336.97	1336.97	4.981148	0.026593			
Residual	229	61464.96	268.4059					
Total	230	62801.93						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	6.338609	1.248512	5.07693	7.94E-07	3.878569	8.798649	3.878569	8.798649
favscor_o_diff	0.265373	0.118903	2.231849	0.026593	0.03109	0.499657	0.03109	0.499657

TWO REGRESSIONS: Top regression is based on the dummy variable while bottom regression is based on the actual difference (fso1-uso1). You'll note that the differential works better than the dummy variable (t-stat is significant, r-squared is higher) because the dummy variable destroys data. If fso1 is only 1 greater than uso1 then that's not nearly as important as fso1 being 10 greater than uso1 but the dummy variable ignores that variation in the data. Don't destroy your data!

B4. Run a regression with the score differential outcome as the dependent variable (Y) and include for the independent variables the Vegas line, "Fans_4_Fav", and the best on-the-field performance measure. What variables are statistically significant? Why do you think the statistical significance changed? How has the R-squared changed from question B1 and why?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.445588							
R Square	0.198548							
Adjusted R Square	0.187956							
Standard Error	14.8906							
Observations	231							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	3	12469.22	4156.407	18.74535	6.73E-11			
Residual	227	50332.71	221.73					
Total	230	62801.93						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.64871	1.602088	-1.0291	0.304527	-4.80558	1.508152	-4.80558	1.508152
line	1.102241	0.161411	6.82878	7.75E-11	0.784185	1.420297	0.784185	1.420297
favscor_o_diff	-0.19921	0.130365	-1.52812	0.127876	-0.45609	0.057667	-0.45609	0.057667
fansforfav	-9E-06	1.78E-05	-0.5046	0.614331	-4.4E-05	2.61E-05	-4.4E-05	2.61E-05

Line is statistically significant but nothing else is here. It's likely that the information in favscor_o_diff is already included in the vegas line. Vegas probably thinks about scoring offense when they make their line. Additionally, the only reason why favscor_o_diff was significant before was because of an omitted variable bias. When the R-squared stat is very low it can be an indicator of omitted variable bias. R-squared stat is basically the same as B1 which suggests that the additional X variables don't do much to predict Y.

Data Mining

C1. Do your best. Forecast the Vegas line using any of the information here (not including any outcome information) and any combination/transformation of the data you desire. **DO YOUR BEST!**

C2. Do your best. Forecast the score differential outcome using any of the information here any combination/transformation of the data you desire. **DO YOUR BEST!**