Homework 1 – College Football Line and Outcomes Database

Data Reading and manipulation

A1. What percentage of games is won by the underdog? What percentage of games is won by the underdog when the favored team is favored by greater than or equal to 10 points? What percentage of games is won by the underdog when the favored team is favored by less than 10 points?

A2. What percentage of favorites have greater rushing offense than the underdog? What percentage of favorites have greater passing offense than the underdog? 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?

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.

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?)

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?

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?

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?

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?

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.

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