Problem Set 1.

Due Feb. 1st 11:59 pm

Thanks to Dr. Michael Lopez, chief of Data Analytics at the NFL, follow on twitter @statsbylopez

In this assignment, you will gain experience implementing regressions to estimate the probability of successful NFL field goals given play and game specific conditions and review basic econometrics. We will also see whom you think is the best kicker in the sport and how one might value them. While field goal kicks are not the most exciting players in the game, the analysis provides an insight into attempting to measure sports ability, which is crucial to the rest of the class. Results from this problem set will be implemented into a model of optimal NFL strategy in later lectures.

Download the dataset kickers2.csv from Canvas.

Explore the data. If any data are strings, generate a version that is a double to be used in regressions. Create a unique indicator for each kicker

1. What was the minimum distance of a field goal kicked in this sample? What was the maximum? Mean? Median?
2. Why isn’t the minimum lower? (For those who are not familiar with football, please read about how field goal distance is measured and its relationship to where the ball is on the field.)
3. What special circumstances might explain the maximum? (Hint: football is a game with 4, 15-minute quarters. At the end of the second quarter there is a halftime break and possession is assigned based on the result of a first-half coin toss)
4. Over the entire sample what percentage of kicks from 40 to 45 yards were made? Kicks over 45 yards?
5. Was the make rate higher on grass or on turf? Is that difference statistically significant? Do you think this is the true effect of surface? Why or why not?[[1]](#footnote-1) (Answer this by doing an OLS regression. For the entire assignment, let’s use the heteroskedasticity robust standard errors, r in stata or the equivalent in R)
6. How is distance of attempt correlated with surface? What might explain this? (Coaches get to choose when to kick a field goal, one is never forced)
7. How is distance correlated with make percentage?
8. What is the formula for omitted variable bias?
9. Given (a) what should happen to the estimate of the effect of a kick being on grass when you add in distance? Verify this is true.
10. Run an ols regression of kick success on distance, surface, point differential, and clock time. Interpret the coefficients. Does it seem like kickers do better or worse late in the game? Does the score of the game seem to effect them?
11. Now add in kicker fixed effects (i.kicker in Stata), what do these correct for? How does adjusted r-squared change?
12. After you run the regression in part 6, do the command to predict fitted values from this regression: “predict, xb” in stata and equivalent in R “predict.lm”. Based on this, what would you predict the probability of Justin Tucker cutting the lead to 8 (scorediff was -11) in 2015, when the gameminute was 30, and he was on turf (row 11083 in the .csv)
13. Does this estimate strike you as reasonable?
14. What would the estimate be for an average kicker?
15. Now run a logistic regression with the same specification as in question 6. Use the command predict. Now what is the predicted probability of Tucker making that field goal? (the predict command in stata is now just “predict”)
16. Why do the coefficients look so different for the logistic regression vs. OLS
17. Who would you say was the best kicker in the NFL over this period? Why? Define best in at least two different ways. Try to quantify the size of the difference.
18. Are these differences stable over time? For example, if players switch team or year over year?
19. Some argue that kickers get better with experience, in this dataset do you see evidence to support this conjecture? Try both a linear and quadratic specification. (For simplicity assume that there were no kicks attempted before the beginning of the dataset). What might be wrong with your estimates (besides incomplete data)? Explain.
20. What are the omitted variables you would want in this dataset? List at least 3 and which direction the bias of excluding them could go and why.

1. Some teams play outdoors on natural grass, others play outdoors on turf, others play indoors on turf. [↑](#footnote-ref-1)