Name:

Stats in practice #7 (optional)

Predicting probability of success

# Field Goals in the NFL

Can we predict the probability that a NFL kicker makes a field goal based on how far away the kicker is from the goal posts? This type of problem is unique, because the response variable away the kicker is from the goal posts? This type of problem is unique, because the response variable *y* is defined as follows

$$ y= \left\{\begin{array}{c}1, success\\0, failure\end{array}\right.$$

This type of data is referred to as “binary” (or sometimes “Bernoulli” in Statistics), because it can only take 2 values. The *x* variable will be a quantitative variable describing the number of yards that the kick is taken from.

# Simple linear regression

## The first question we might have is “why not just use simple linear regression”? We explore that here, and show why it doesn’t make sense for this type of data.

## Loading the data.

### Load the data into R by typing:

### nfl\_data <- read.table('http://users.stat.ufl.edu/~winner/data/fieldgoal.dat')

### distance <- nfl\_data$V1

### x <- nfl\_data$V2

## Linear regression

### Consider the simple linear regression equation:

### $$y=a+bx$$

### where y is the response variable (1 if make, 0 if miss) and x is the distance. Find the slope and intercept in R by typing:

### fit\_slr <- lm(y~distance)

### Type “fit\_slr” into the console. Write down the least squares regression line equation below:

### Use your regression line to estimate the probability that a kick is successful from *x=55* yards. Does this seem reasonable?

### Use your regression line to estimate the probability that a kick is successful from *x=18* yards. Does this make any sense? What went wrong?

### Make a scatterplot of the data and add the regression line by typing:

### plot(distance, Y, xlim=c(0, 70))

### abline(fit\_slr, col=’red’)

### **Copy and paste this plot below.**

# Predicting probability of success with Logistic Regression

## The main problem with using logistic regression for probabilities, is that we have no way of ensuring that the prediction is a valid probability (i.e. between 0 and 1). Logistic regression gives us one possible way of fixing this issue.

## Logistic Regression

### In Logistic regression, we model the probability directly with a function that must always output values between 0 and 1.

### $$Probability of success=\frac{1}{1+e^{-(a+bx)}}$$

### This looks more complicated, but we can use this in a very similar way to predict probability of success. This model can be fit in R by typing:

### fit\_slr <- glm(y~distance, family=binomial)

### Type “fit\_slr” into the console. What are the estimated values of *a* and *b?* Write out the logistic regression equation for predicting success.

### Use your logistic regression equation to predict the probability that a kick is successful from *x=55* yards. How does this compare to what you got with simple linear regression?

### Use your logistic regression equation to predict the probability that a kick is successful from *x=18* yards. Explain why this is better than the simple linear regression answer.

### Make a scatterplot of the data and add the regression polynomial with the following code. (You will have to replace the “?” symbols with your estimated values of *a* and *b*.

### plot(distance, Y, xlim=c(0, 70))

### curve(1/(1+exp(-(?+?\*x))), add=TRUE, col=’red’)

### **Copy and paste this plot on the following page**