What is the Predicted Probability of a Field Goal Given Yards?

Chirag Kashyap

Introduction

In football, offenses have four downs to travel ten yards or they turn the ball over. So on fourth down, teams have a decision to go for it, kick a field goal, or punt. Depending upon the situation, coaches make their decisions and have to live with them. In some circumstances, the ability to make a field goal from a certain distance can make or break a team's season. Therefore, it is important to know the probability of making a field goal at a certain distance in order to make correct decisions to win the game. In this project, I will find the predicted probability of making a field goal from 18 to 62 yards, according to our dataset.

Material and Methods

The data used in this project was from the 2003 NFL regular season. All 948 field goal attempts from all 17 weeks of the regular season were used and are depicted below. This dataset was found on http://www.stat.ufl.edu/~winner/datasets.html and the data itself was sourced from www.espn.com and www.jt-sw.com.

	Yards								
Outcome	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
Made	1	5	18	17	36	33	28	17	26
Missed	0	0	1	1	0	0	2	0	1
	27	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
Made	32	31	33	19	23	16	30	19	20
Missed	0	1	4	3	3	2	3	6	7
	36	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**
Made	21	26	26	26	24	22	26	21	20
Missed	4	6	7	7	5	10	2	7	11
	45	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**
Made	13	25	20	21	16	14	12	5	7
Missed	8	9	7	15	13	11	6	9	7
	54	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**
Made	3	1	1	1	1	0	0	0	0
Missed	5	5	0	1	0	0	2	0	1

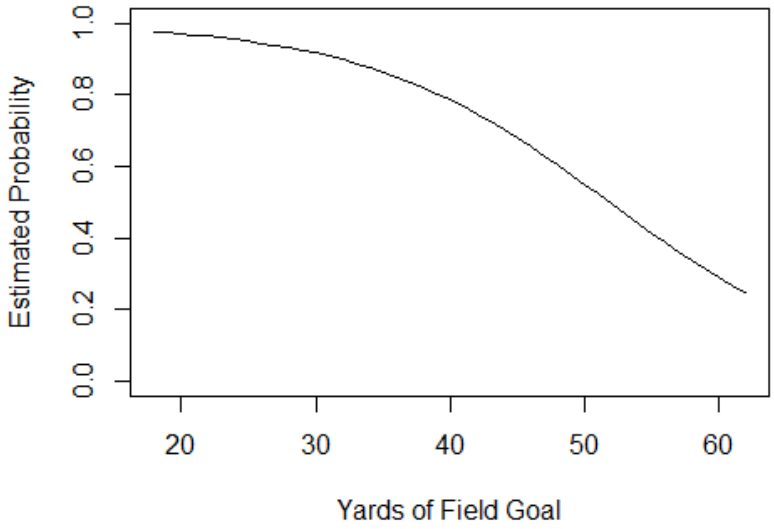
There are three variables in this data set: yards, outcome, and week. For this project, I only used two: yards and outcome. Yards indicates the distance of the field goal in yards. Because yards has a large number of values, it can be treated as a continuous variable. Outcome is our binary response variable which indicates if the kicker made (1) or missed (0) the field goal. Using simple logistic regression, I can model the probability of a kicker making a field goal with the following formula:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

I will estimate the parameters β_0, β_1 utilizing simple linear regression and use Wald hypothesis test and Wald confidence intervals to make sure they are statistically significant. Percent concordant and percent discordant will be used to analyze association between observed and predicted probabilities. I will also analyze the standardized Pearson residuals to make sure none of them deviate. Finally, I will display the predicted probability of making a field goal from 18 yards to 62 yards.

Results

The probability of a kicker making a field goal can be modeled with the maximum likelihood estimates for β_0 and β_1. Using simple logistic regression, I found that $\beta_0 = 5.6979$ and $\beta_1 = -0.1099$. Thus, the model is $\pi(x) = (e^{5.6979 - 0.1099x})/ (1 + e^{5.6979 - 0.1099x})$. The model tell us that the predicted probability of making a field goal decreases the longer the field goal is. This is seen from the predicted probability plot below.




```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.69788    0.45110   12.63   <2e-16 ***
yards       -0.10991    0.01058  -10.38   <2e-16 ***
```

Now we test β_0 and β_1 using the Wald hypothesis test, with $H_o : \beta = 0$ and $H_a : \beta \neq 0$. For the intercept (β_0), we have a z-value of 12.63 and a p-value of <0.0001, so we reject H_o and conclude H_a at an $\alpha = 0.05$. For β_1, we have a z-value of -10.36 and a p-value of <0.0001, so we reject H_o and conclude H_a at an $\alpha = 0.05$. Likewise, the 95% Wald confidence interval for β_0 is $5.6788 \pm (1.96)0.4511 = (4.7946, 6.5630)$ and the 95% Wald confidence interval for β_1 is $-0.1099 \pm (1.96)0.0106 = (-0.1307, -0.0891)$. Both of the confidence intervals agree with my conclusion

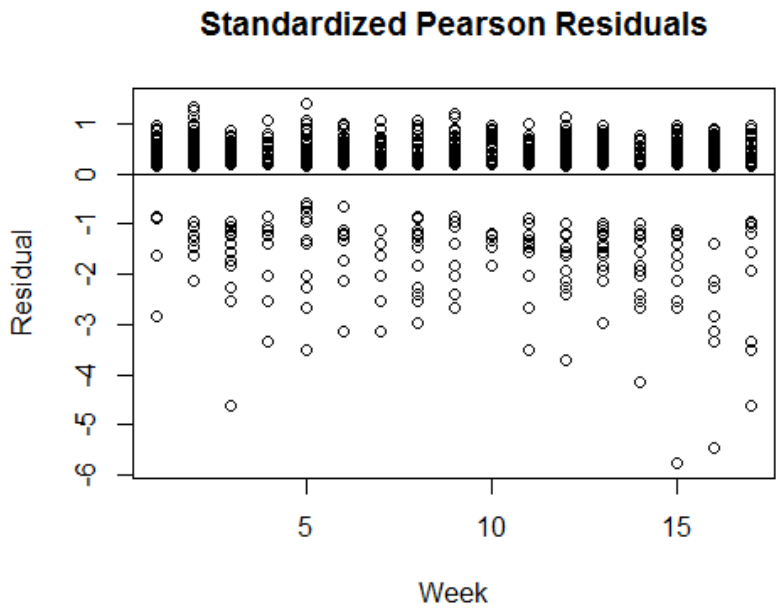
to reject the null hypothesis.

The 95% confidence interval of the odds ratio can be determined by taking the exponential of β_1: ($e^{-0.1307}$, $e^{-0.0891}$) or (0.8775, 0.9148). So when the field goal is increased in distance by another yard, we are 95% confident that the odds of making the field goal will decrease by 8.52% to 12.25%. The confidence interval shows that our predicted probability curve will be monotonic decreasing as seen above.

A concordance percentage of 74.95% indicates a moderate association between the predicted and observed probabilities of success. The data has 756 field goals made out of 948 attempts. Therefore the model should be satisfactory in predicting made field goals.

Percent Concordant	Percent Discordant	Percent Tied	Number of Pairs
74.95%	22.65%	2.40%	145152

The standardized Pearson residuals of the field goals are ordered on which week the game was played, in order to see if there was a trend over time. There does not seem to be a trend over time, but there are a large number of residuals less than -2. All of these residuals less than -2 are short yardage field goals that have missed. It appears that while this model does a good job of predicting makes, it does not do a good job of predicting and accounting for misses. Because of the residual plot, there is a serious lack of fit regarding predicting missed field goals.



The predicted probabilities of making a field goal from 18 yards to 62 yards is below. Between 51 and 52 yards is when the probability of making a field goal is exactly 50%. Coaches should take this into account when making their decision about what to do on fourth down. They might be better off punting when making a field goal is not the predicted outcome.

Yards	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
Probability	0.9763	0.9736	0.9707	0.9674	0.9637	0.9597	0.9552	0.9503	0.9448
	27	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
Probability	0.9388	0.9322	0.9249	0.9169	0.9081	0.8985	0.8880	0.8766	0.8642
	36	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**
Probability	0.8508	0.8363	0.8207	0.8040	0.7861	0.7670	0.7468	0.7255	0.7030
	45	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**
Probability	0.6796	0.6552	0.6300	0.6040	0.5774	0.5504	0.5231	0.4956	0.4682
	54	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**
Probability	0.4410	0.4141	0.3877	0.3619	0.3370	0.3129	0.2897	0.2676	0.2467

Conclusion and Discussion

Using field goal outcome data from the 2003 NFL regular season, I fit a simple logistic regression model. Both β_0 and β_1 were found to be significant by both the Wald hypothesis test and Wald confidence interval. As seen in the predicted probability plot, we have a monotone decreasing curve, so the longer the distance of the field goal, the less likely the kicker is too make it. The 95% confidence interval of the odds ratio also shows this same thing. The concordance percentage indicates a moderate association between the predicted and observed probabilities of success. However, the standardized Pearson residuals show massive deviation because of the short-range missed field goals in the data. This means that this model is not good at predicting missed field goals. A multiple linear regression with more situational data could help with this problem, since many of these missed field goals happen in high-pressure, late-game situations. That would help the model take into account why or how these field goals were missed. Nonetheless, I was surprised to see so many of the missed field goal residuals be below -2. Lastly, I predicted the probabilities of all field goals taken between 18 yards and 62 yards. This analysis would be useful to coaches when they are decided what to do on fourth down. In a late game situation, missing a field goal could lead to the opposing team taking advantage of the field position to score a touchdown and effectively end the game. Sometimes playing it safe is the way to go.

Code Appendix

```
#STA 138 Project 2
#Chirag Kashyap 998388067
library(boot)

nfl <- read.table("C:/Users/ckashyap/Desktop/Current Courses/STA 138/Datasets/nfl.txt", quote="\"",
comment.char="")
names(nfl) = c("yards", "outcome", "week")
table(nfl[,1:2])

fgyard = glm(outcome~yards,data=nfl,family=binomial(link=logit))
```

```
summary(fgyard)
prob_fgyard=predict(fgyard,data.frame(yards=seq(18,62)),type="response")
plot(seq(18,62),prob_fgyard,xlab="Yards of Field Goal",ylab="Estimated Probability",
   ylim=c(0,1),type="l")
plot(nfl$week, glm.diag(fgyard)$rp, ylab = "Residual", xlab = "Week",
   main = "Standardized Pearson Residuals" )
abline(0,0)
concordance(fgyard)

#from https://discuss.analyticsvidhya.com/t/how-to-get-the-percentage-concordant-and-discordant-values-for-a-logistic-regression-model-in-r/1458/2
concordance<-function(model){
 # Get all actual observations and their fitted values into a frame
 fitted<-data.frame(cbind(model$y,model$fitted.values))
 colnames(fitted)<-c('respvar','score')
 # Subset only ones
 ones<-fitted[fitted[,1]==1,]
 # Subset only zeros
 zeros<-fitted[fitted[,1]==0,]

 # Initialise all the values
 pairs_tested<-0
 conc<-0
 disc<-0
 ties<-0

 # Get the values in a for-loop
 for(i in 1:nrow(ones)){
  for(j in 1:nrow(zeros)){
   pairs_tested<-pairs_tested+1
   if(ones[i,2]>zeros[j,2]) {conc<-conc+1}
   else if(ones[i,2]==zeros[j,2]){ties<-ties+1}
   else {disc<-disc+1}}}
 # Calculate concordance, discordance and ties
 concordance<-conc/pairs_tested
 discordance<-disc/pairs_tested
 ties_perc<-ties/pairs_tested
 return(list("Concordance"=concordance,
         "Discordance"=discordance,
         "Tied"=ties_perc,
         "Pairs"=pairs_tested))}
```