UNITED STATES MILITARY ACADEMY

SCORE Project Presentation

MA388: SABERMETRICS

SECTION D1

LTC MICHAEL POWELL

BY

CDT MICHAEL KRAMER '24, CO G4

CDT RYAN KROLIKOWSKI '24, CO G4

WEST POINT, NEW YORK

05 MAY 2024

RK
_____ WE CERTIFY THAT WE HAVE COMPLETELY DOCUMENTED ALL SOURCES THAT WE USED TO COMPLETE THIS ASSIGNMENT AND THAT WE ACKNOWLEDGED ALL ASSISTANCE I RECEIVED IN THE COMPLETION OF THIS ASSIGNMENT.

_____ WE CERTIFY THAT WE DID NOT USE ANY SOURCES OR RECEIVE ANY ASSISTANCE REQUIRING DOCUMENTATION WHILE COMPLETING THIS ASSIGNMENT.

SIGNATURES: Ryan Krolikowski

SCORE Project

Fourth Down Football and Logistic Regression

CDT Ryan Krolikowski & CDT Michael Kramer

May 5th, 2024

Learning Goals:

- Learn and apply basic functions in R with NFL data.
- Learn and apply a Logistic Regression model in R.
- Interpret coefficients and their meaning from a Logistic Regression model.

Introduction

Have you ever wondered what goes through a coaches head when they go for it on fourth down from their own 30 yard line with four minutes left in the fourth quarter? Well today we are going to dig deeper in the probabilities behind the decisions that coaches make and we can find out using statistics to determine whether or not that was a good idea. Today, we will use a logistic regression model can we determine whether to go for it on 4th down based on field position and yards to go from first down line. Today we will address the following questions:

- What is a Logistic Regression Model?
- How do I interpret the coefficients?

Logistic Regression is a statistical method used for binary classification and allows us to analyze and predict outcomes that have two possible values in this case go for it or punt. By applying this technique to football, we can examine the factors that influence a coach's decision to go for it on fourth down, such as field position and the yards remaining to secure a first down. Unlike linear regression, which predicts continuous outcomes, logistic regressions models the probability that a given input belongs to a given category.

Here is the logistic regression equation:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Where:

- p is the probability of the dependent variable equaling a case (coded as 1), given the independent variables.
- $\frac{p}{1-p}$ is the odds ratio, representing the odds of the event occurring (probability of the event divided by the probability of the event not occurring).
- $\log\left(\frac{p}{1-p}\right)$ is the natural log of the odds ratio, known as the log-odds or the logit function.
- β_0 is the intercept from the regression equation, representing the log-odds of the dependent variable being 1 when all the independent variables are 0.

- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables $X_1, X_2, \ldots, X_n$. Each coefficient represents the change in the log-odds of the dependent variable being 1 for a one-unit change in the corresponding independent variable, holding all other variables constant.

To obtain the predicted probability (p) from the log-odds, you can use the logistic function, which is the inverse of the logit function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

This equation allows you to calculate the probability p that the dependent variable is 1 given the values of the independent variables $(X_1, X_2, \ldots, X_n)$, incorporating the estimated coefficients $(\beta_0, \beta_1, \beta_2, \ldots, \beta_n)$ from the logistic regression model.

Today, we will go embark on our statistical exploration to uncover the underlying coaching decisions on fourth down using logistic regression to answer the pressing question:

- Can we accurately predict the likelihood of a team opting to go for it on fourth down based on their field position and the yards to go?

Data

In order to tackle our problem we will use NFL play-by-play data from the 2023 football season from the NFL fast R data set. The data contained in the data frame consists of various statistics relevant to the given play including the number of timeouts a team has, their position on the field, whether or not a touchdown was scored, as well as other categories for the outcome of the play. The data frame has 49,665 plays from the start of the regular season to the Superbowl and 372 separate variables to describe the situation and outcome of a given play.

- Load the necessary packages:

```
library(nflfastR)
library(tidyverse)
library(kableExtra)
library(modelsummary)
library(pROC)
```

- Create our dataset. Take a look at the structure of the dataset using the kable display package:

```
data <- load_pbp(2023)

data %>%
  select(1:7) %>%
  head(2) %>%
  kable() %>%
  kable_classic()
```

play_id	game_id	old_game_id	home_team	away_team	season_type	week
1	2023_01_ARI_WAS	2023091007	WAS	ARI	REG	1

39	2023_01_ARI_WAS	2023091007	WAS	ARI	REG	1

Methods/Instructional Content:

- Logistic Regression - UC Business Analytics R Programming Guide (https://uc-r.github.io/logistic_regression). This Guide provides a comprehensive look at logistic regression models in R and all about how they work in R and how to interpret the coefficients. It provided me with a great background on the the model how I could apply it to this data set.
- An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (https://www.statlearning.com/). This textbook is a great resource for application of Logistic Regression in R and learning about more advanced classification models in R.I learned more about interpreting the coefficients and how to make predictions using a logistic regression which helped me to formulate my research question.

Exercises/Activities:

Lets now answer our original question:
What is the probability of attempting a fourth down conversion given a team's field position?

1. *Formulate a list of relevant variables to answering the question and filter the data to those variables.* What variables can you think of being relevant to the success and failure of a fourth-down conversion? Our analysis brings us to considering distance from endzone, distance from the first-down marker, the score differential, and the offensive team's win probability. For simplicity, we will use distance from the endzone and distance to the first-down marker. Let's clean up our dataset using those variables.

```
fourth_down_plays <- data %>%
  filter(down == 4) %>%
  mutate(go_for_it = ifelse(play_type == "no_play" & penalty == TRUE, NA, play_type %in% c("pass", "run


fourth_down_plays <- fourth_down_plays %>%
  select(yardline_100, ydstogo, go_for_it)

fourth_down_plays %>%
  head(5) %>%
  kable() %>%
  kable_classic()
```

yardline_100	ydstogo	go_for_it
49	7	FALSE
70	8	FALSE
11	9	FALSE
36	9	FALSE
85	11	FALSE

2. *Apply the logistic Regression Model in R.*
The logistic regression model in our context looks like this:

$$ProbabilityOfGoingForIt = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 FieldPosition_i + \beta_2 YardsToGo_i$$

	(1)
(Intercept)	0.645
	(0.098)
yardline_100	−0.017
	(0.002)
ydstogo	−0.213
	(0.012)
Num.Obs.	4229
AIC	3550.1
BIC	3569.1
Log.Lik.	−1772.029
RMSE	0.36

We must apply this function in R following the below format:

$$model < -glm(outcome\ predictor1 + predictor2, family = binomial(link = "logit"), data = data)$$

This equation fits a logistic regression model predicting outcome from predictor1 and predictor2 within the data.

```
# Fit the logistic regression model
model <- glm(go_for_it ~ yardline_100 + ydstogo, family = binomial(link = "logit"), data = fourth_down_

# Summary of the model
model %>%
  modelsummary()
```

3. *Interpret the coefficients and write a succinct summary of your findings.*

(Intercept) 0.645: The intercept in a logistic regression model represents the log odds of the dependent variable (deciding to go for it on fourth down) being 1 (yes) when all the independent variables are 0.

yardline_100 -0.017: The coefficient for yardline_100 is negative, indicating that as the team gets closer to the opponent's end zone (a decrease in yardline_100), the likelihood of deciding to go for it on fourth down increases. For each yard closer to the opponent's end zone, the log odds of going for it decrease by approximately 0.017, implying higher odds of going for it.

ydstogo -0.213: The coefficient for ydstogo is also negative, showing that as the yards to go for a first down increase, the likelihood of opting to go for it decreases. This coefficient indicates a relatively stronger effect compared to yardline_100, with each additional yard to go decreasing the log odds of going for it by about 0.213.

4. *Measure the models ability to discriminate between the decision to go for it or not (Hint: AUC).* The Area Under the Curve (AUC) is a metric used to evaluate the performance of binary classification models, such as logistic regression. The AUC measures the ability of the model to correctly classify the positive and negative classes across all possible threshold values. The AUC is part of the Receiver Operating Characteristic (ROC) curve analysis. Here's how the AUC is conceptually described in text:

ROC Curve The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR, also known as sensitivity) against the false positive rate (FPR, 1 - specificity), at various threshold settings.

AUC The Area Under the ROC Curve (AUC) represents the measure of the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). Here's what the AUC signifies:

- AUC = 0.5: This value represents a model with no discrimination ability, akin to random guessing.
- 0.5 < AUC < 1.0: Represents a model with some ability to discriminate between the positive and negative classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.
- AUC = 1.0: This value indicates a perfect model that can perfectly distinguish between all the positive and negative instances.

Here is Base Code For ROC curves in R.

Install and load the pROC package

install.packages("pROC") library(pROC)

Assume 'predictions' are the probabilities returned by a logistic regression model and 'actual' are the actual binary outcomes (0 or 1)

roc_curve <- roc(actual, predictions)

Calculate the AUC

auc_value <- auc(roc_curve) print(auc_value)

- Calculate predicted probabilities using the logistic regression model. *type = "response"* specifies that the output should be the probability estimates.

```
probabilities <- predict(model, newdata = fourth_down_plays, type = "response")
```

- Convert the *go_for_it* column from a logical to a numeric type, and subtract 1 to make it binary (0 or 1).

```
actual_outcomes <- as.numeric(fourth_down_plays$go_for_it) - 1
```

- Calculate the ROC curve using the actual outcomes and the predicted probabilities.

```
roc_curve <- roc(actual_outcomes, probabilities)
```

- Calculate the area under the curve (AUC) from the ROC curve.

```
auc_value <- auc(roc_curve)
```

- Display the AUC value.

```
auc_value
```

```
## Area under the curve: 0.7892
```

- **What can you interpret about the model from the AUC value we found?**
- **What are potential ways we can increase model accuracy on the decision of whether or not to go for it on fourth down?**
- **Can we increase model accuracy by increasing the number of variables that we look at in our logistic regression since we currently only have two variables to assess the impact on going for it on fourth down?**

Wrap-Up/Conclusions

The use of logistic regression in sports analytics especially in football highlights the powerful intersection between statistical techniques and sports strategy, offering a framework for making more informed decisions and uncovering deeper insights into game dynamics. Our analysis confirmed that teams are more likely to go for it on fourth down when they are closer to the opponent's end zone and when they have fewer yards to go for a first down.This can be valuable for defensive teams to get either their defensive team ready or punt team ready for a given fourth down.

Work Cited

"An R Package to Quickly Obtain Clean and Tidy NFL Play by Play Data." *Www.nflfastr.com*, www.nflfastr.com/.

James, Gareth, et al. An Introduction to Statistical Learning : With Applications in R. SECOND ed., Springer, 2013.

"Logistic Regression · UC Business Analytics R Programming Guide." *Github.io*, 2019, uc r.github.io/logistic_regression.

OpenAI. "Provide the logistic regression equation written out with variables explained" ChatGPT, OpenAI, 12 April 2024.

OpenAI. "help formulate AUC code and background as we wanted a way to depict the models ability to discriminate whether or not to go for it on fourth down" ChatGPT, OpenAI, 12 April 2024.