Logistic Regression - Introduction

Ryan Miller

Modeling Binary Outcomes

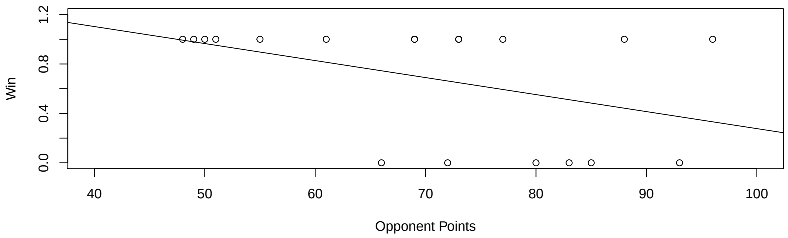
- At this point in the semester, we've spent several weeks modeling numerical outcomes
    - Unfortunately, these models aren't suitable for categorical outcomes
- This week, we'll introduce **logistic regression**, which is perhaps the most widely used model for binary categorical outcomes

Introduction

- Consider the XU basketball team dataset, we might be interested in the outcomes "win" and "loss"
  - If we create a dummy variable that encodes a numeric value of "1" to a "win" and "0" to a "loss", we interpret $E(y)$ as the *probability of a win*

Introduction

- Consider the XU basketball team dataset, we might be interested in the outcomes "win" and "loss"
  - If we create a dummy variable that encodes a numeric value of "1" to a "win" and "0" to a "loss", we interpret $E(y)$ as the *probability of a win*
- The graph below shows the simple linear regression model: `Win ~ OppPts`
  - What problem does this model exhibit?



Generalized Linear Models

Logistic regression, and linear regression, are both types of **generalized linear models** (GLMs for short):

$$g(E(y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

All GLMs have the following components:

- *Systematic component* - a linear combination of predictor variables (ie: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$)
- *Random component* - a probability distribution for Y, the outcome variable
- *Link function* - a function that links $E(Y)$ to the model's systematic component

Link Functions

- In linear regression, the link function is simply the identity function (ie: $g(E(y)) = E(y)$)
  - This link makes interpretations very straightforward, as predictors influence the outcome directly

Link Functions

- In linear regression, the link function is simply the identity function (ie: $g(E(y)) = E(y)$)
  - This link makes interpretations very straightforward, as predictors influence the outcome directly
- For binary outcomes, it would be unwise to use an identity link function
  - In these situations, $E(y)$ can been seen as a *probability*
  - Thus, any model should be careful to avoid generating predictions for $E(y)$ that are outside $[0, 1]$

Odds

- An alternative way of expressing the likelihood of an event is the *odds of the event*
  - The *odds* of an event is a ratio of how often the event occurs relative to how often it does not occur
- If an event has a 50% probability, the odds are 1, which are often called "1 to 1 odds"
- If an event has a 75% probability, the odds are 3, which are often called "3 to 1 odds"

Odds

- The odds of an event can range from 0 (ie: $0/1$) to $+\infty$ (ie: $1/0$)
  - This makes odds a more desirable modeling outcome than probability
  - The chances of a linear combination of predictors resulting in a value outside $[0, \infty]$ is lower than getting a prediction outside of $[0, 1]$

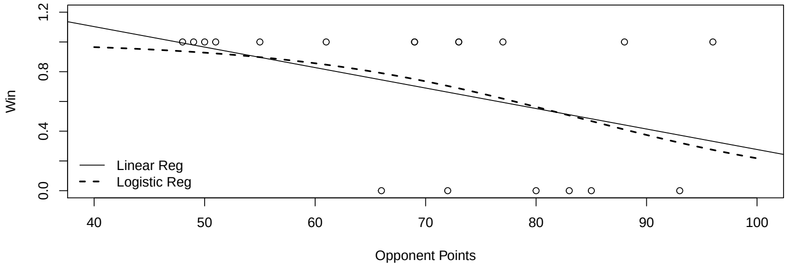
Odds

- The odds of an event can range from 0 (ie: $0/1$) to $+\infty$ (ie: $1/0$)
  - This makes odds a more desirable modeling outcome than probability
  - The chances of a linear combination of predictors resulting in a value outside $[0, \infty]$ is lower than getting a prediction outside of $[0, 1]$
- Further, the *log-odds*, or *logit*, of an outcome (ie: $ln(\frac{E(y)}{1-E(y)})$ can take-on values ranging from $-\infty$ to $+\infty$
  - Logistic regression uses a *logit* link function

The Logistic Regression Model

Logistic regression uses the *logit* link function, the *binomial* probability distribution, and a linear combination of predictors:

$$log\left(\frac{E(y)}{1-E(y)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$



Interpretting Model Coefficients

- In logistic regression, each predictor makes a linear contribution towards the log-odds of an event
  - Thus, each additional point scored by Xavier's opponent is expected to decrease the log-odds of winning by 0.077
  - Unfortunately, the log-odds scale is not very easily interpreted

```
ml <- glm(win ~ Opp.1, data = xub, family = "binomial")
ml$coefficients
```

```
## (Intercept)       Opp.1
##  6.38825140 -0.07667347
```

Interpretting Model Coefficients

Fortunately, we can use mathematics to make sense of things:

$$log(\frac{E(y)}{1-E(y)}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$
$$\implies \frac{E(y)}{1-E(y)} = exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)$$
$$\implies \frac{E(y)}{1-E(y)} = exp(\beta_0) * exp(\beta_1 X_1) * \ldots * exp(\beta_p X_p)$$

- The exponent of the intercept represents the *baseline odds*
- The exponent of $\beta_1, \ldots, \beta_p$ is a *multiplier* of the baseline odds

Interpretting Model Coefficients

- In our XU basketball example, $exp(6.39) = 595.9$ and $exp(-0.077) = 0.92$
  - The baseline odds reflect the likelihood of XU winning if the opponent doesn't score (somewhat meaningless)
  - Then, we estimate an 8% decrease in the odds of XU winning for each point scored by the opponent

Odds Ratios

- For binary predictors, *exp*(β) yields an *odds ratio*

```
ml <- glm(win ~ Location, data = xub, family = "binomial")
exp(ml$coefficients)
```

```
## (Intercept)   LocationH
##         0.5        11.0
```

- The odds XU winning at home are 11 times the odds of XU winning on the road
  - We can verify this with a *contingency table*, odds at home = 11/2, odds on the road = 2/4, odds ratio = $(\frac{11/2}{2/4} = 11)$

```
##
##      Loss Win
##    A    4   2
##    H    2  11
```

Closing Remarks

- Logistic regression is a popular modeling approach because it yields sensible, interpretable models that can be used for statistical inference on binary outcomes
- Unfortunately, at least in some aspects, is that logistic regression focuses on odds rather than probabilities
  - Odds reflect the likelihood of how often an outcome occurs relative to how often it does not occur
  - The estimated coefficients in Logistic regression, after transformation, can be used to assess the adjusted effect of a predictor on the odds of an outcome