Logistic regression

Applied to estimating tennis match win odds

Dr Emma Davis

9th Nov 2023

Outline

What is logistic regression?

- Introduction
- Simple case: 1 predictor
- The full model
- Odds and interpreting coefficients

Applied example: Predicting WTA match outcomes in R

Summary

What is logistic regression?

Consider a scenario where we have a **binary** variable

$$Y = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases}$$

based on n predictors $X_1, \ldots, X_n$

What is logistic regression?

Consider a scenario where we have a **binary** variable

E.g., the outcome of a tennis match

$$Y = \begin{cases} 1 & \text{if win} \\ 0 & \text{if lose} \end{cases}$$

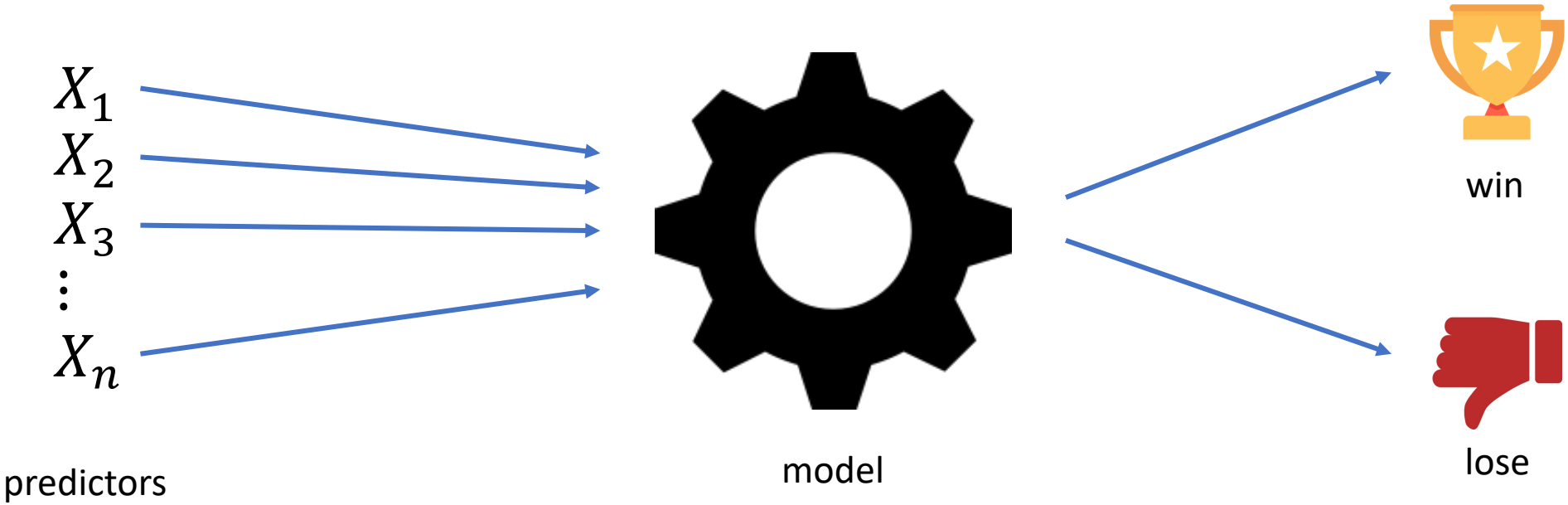
based on n predictors $X_1, \ldots, X_n$

E.g., the difference between your world ranking and your opponent's

What is logistic regression?

Consider a scenario where we have a **binary** variable

E.g., the outcome of a tennis match



What is logistic regression?

Consider a scenario where we have a **binary** variable

E.g., the outcome of a tennis match

Goal: Estimate the probability of observing outcome, Y, for a given set of predictors, X

$$\pi = \mathbb{P}[Y \,|X = (x_1, \ldots, x_2)]$$

What is logistic regression?

$Y \in \{0,1\}$ – indicator of match win success

π – probability of winning the match

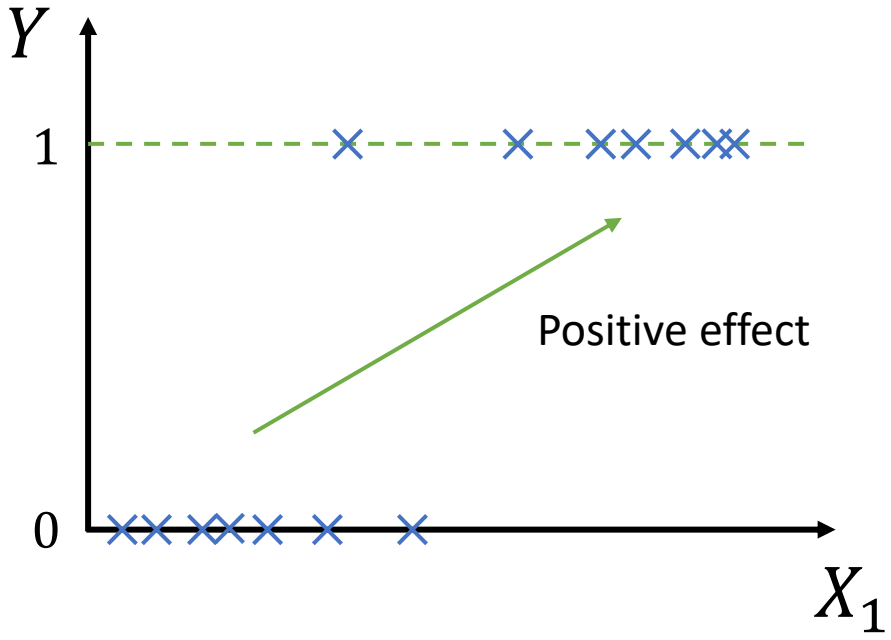
$$\mathrm{Y} \mid \pi \sim \mathrm{Bernoulli}(\pi)$$

With expected value

$$E(Y|\pi) = \pi$$

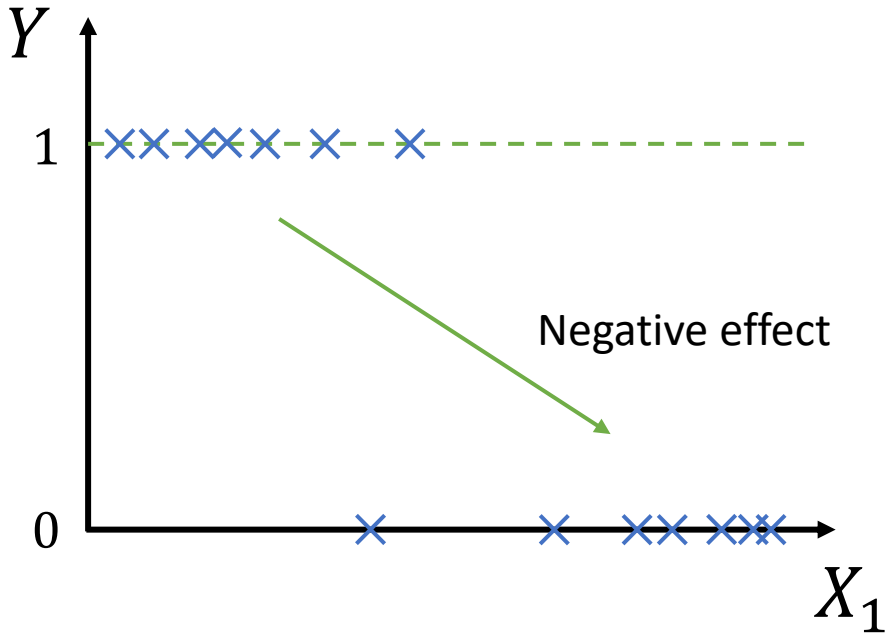
What is logistic regression?

Simple case: 1 predictor - X_1



What is logistic regression?

Simple case: 1 predictor - X_1



What is logistic regression?

Simple case: 1 predictor - X_1

Recall: Linear regression (continuous Y)

$$\pi = \beta_0 + \beta X_1$$

expected value $\in \mathbb{R}$

intercept

slope

What is logistic regression?

Simple case: 1 predictor - X_1

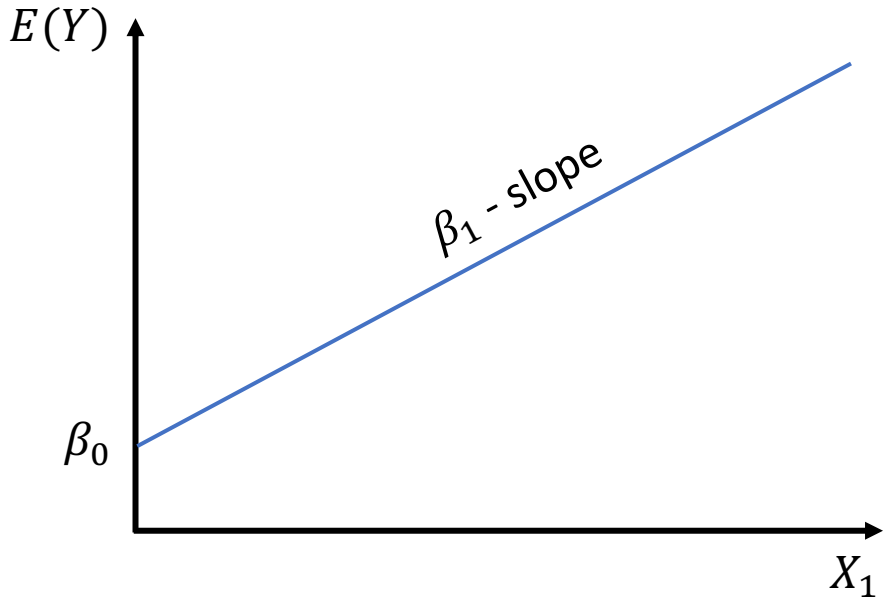
Recall: Linear regression (continuous Y)

$$\pi = \beta_0 + \beta X_1$$

expected value $\in \mathbb{R}$

intercept

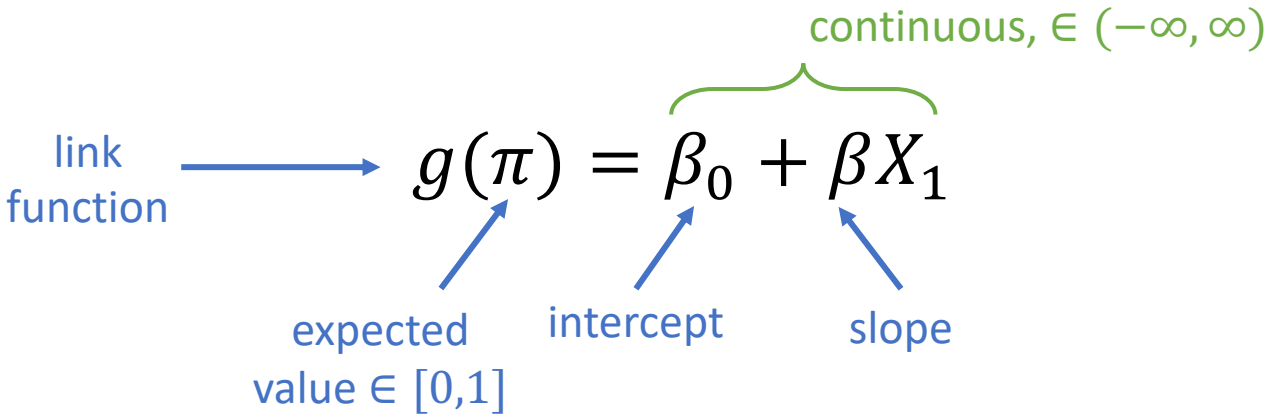
slope



What is logistic regression?

Simple case: 1 predictor - X_1

Recall: Generalised Linear Models



$Y \in \{0, 1\} \Rightarrow \pi \in [0, 1]$

What is logistic regression?

Simple case: 1 predictor - X_1

Recall: Generalised Linear Models

$$\pi = g^{-1}(\beta_0 + \beta X_1)$$

continuous, $\in (-\infty, \infty)$

$\in [0, 1]$

$Y \in \{0, 1\} \Rightarrow \pi \in [0, 1]$

What is logistic regression?

Simple case: 1 predictor - X_1

We need:

- π monotonic with X_1
- $\pi \in [0, 1]$

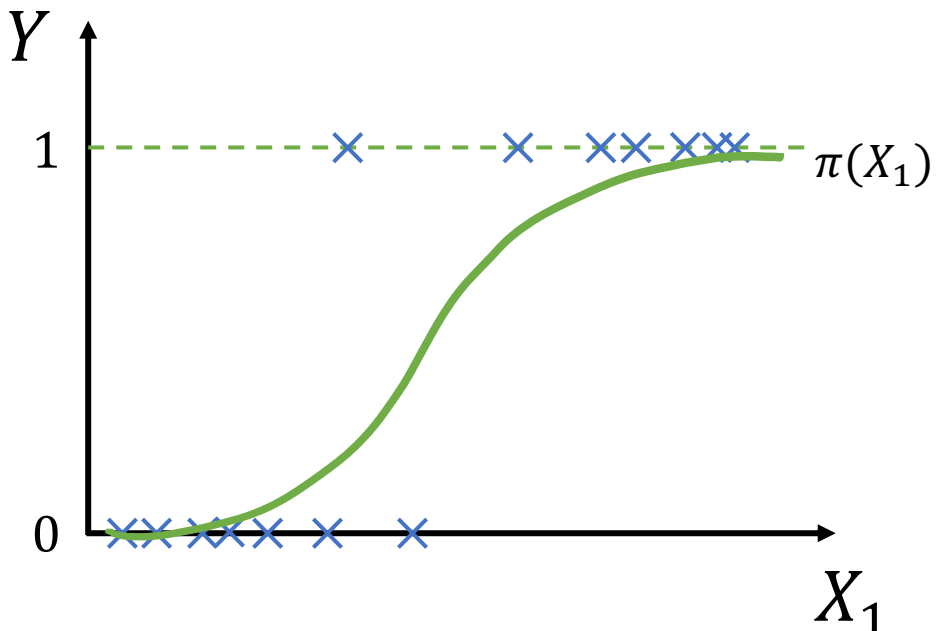
Notice: The logistic function $P(x) = \frac{1}{1+e^{-x}}$ has $P: \mathbb{R} \to [0, 1]$

What is logistic regression?

Simple case: 1 predictor - X_1

So, let's try $g^{-1} = P$

$$\pi(\beta_0 + \beta_1 X_1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}$$



$$P(x) = \frac{1}{1 + e^{-x}}$$

What is logistic regression?

Simple case: 1 predictor - X_1

So, let's try $g^{-1} = P$

$$\pi(\beta_0 + \beta_1 X_1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}$$

This rearranges:

$$\underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{g(\pi) = \text{“log-odds”}} = \beta_0 + \beta_1 X$$

What is logistic regression?

For n predictors, we have the full logistic regression model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \cdots + \beta_n X_n$$

What is logistic regression?

A quick note on odds:

Ratio between prob win and prob lose

$$\text{Odds} = \frac{\pi}{1-\pi} \quad \in [0, \infty)$$

E.g. Match-win probability of $\pi = 0.25$

$$\text{Odds} = \frac{0.25}{1-0.25} = \frac{1}{3}$$

i.e. you are $1/3$ as likely to win as to lose

What is logistic regression?

A quick note on odds:

Ratio between prob win and prob lose

$$\text{Odds} = \frac{\pi}{1-\pi} \quad \in [0, \infty)$$

E.g. Match-win probability of $\pi = 0.5$

$$\text{Odds} = \frac{0.5}{1-0.5} = 1$$

i.e. you are just as likely to win as to lose

What is logistic regression?

A quick note on odds:

Ratio between prob win and prob lose

$$\text{Odds} = \frac{\pi}{1-\pi} \quad \in [0, \infty)$$

E.g. Match-win probability of $\pi = 0.75$

$$\text{Odds} = \frac{0.75}{1-0.75} = 3$$

i.e. you are $3\times$ as likely to win as to lose

What is logistic regression?

A quick note on odds:

Ratio between prob win and prob lose

$$\text{Odds} = \frac{\pi}{1-\pi} \quad \in [0, \infty)$$

E.g. Match-win probability of $\pi = 0.75$

$$\text{Odds} = \frac{0.75}{1-0.75} = 3$$

i.e. you are $3\times$ as likely to win as to lose

Note: Not to be confused with betting odds, where decimal odds $= \frac{1}{\pi} \in [1, \infty)$

What is logistic regression?

Interpreting logistic regression coefficients:

β_0 is the log odds if all of $X_1, \ldots, X_n = 0$

E.g., e^{β_0} = the odds of beating an opponent with the same rank as yourself

β_1 is the typical change in log odds when X_1 increases by 1

(controlling for all other $X_2, \ldots, X_n$)

E.g. if your rank increased by 1 in comparison to your opponent

Example: Women's Tennis

Data taken from: www.tennis-data.co.uk

Includes:

- 23,228 WTA matches from the past 10 years (2014-2023)
- World rank, ranking points (for winner and loser)
- Tournament
- Surface
- Round
- Best available **decimal** betting odds at the time ($= 1/\pi$)

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Example: Women's Tennis

Cleaned and split data into

- Fitting data: 90% of matches
- Testing data: 10% of matches

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Example: Women's Tennis

Player A vs. Player B

Question: what is the probability Player A wins?

Possible predictors:

- X_1 = Difference between WTA rank of Player A and Player B
  Negative ⇒ Player A higher ranked
- X_2 = Difference between WTA ranking points of Player A and Player B
  Positive ⇒ Player A higher ranked (more points)

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Player A vs. Player B
Question: what is the probability Player A wins?

Possible predictors:
- X_1 = Difference between WTA rank of Player A and Player B
  Negative ⇒ Player A higher ranked
- X_2 = Difference between WTA ranking points of Player A and Player B
  Positive ⇒ Player A higher ranked (more points)

Example: Women's Tennis

The model

π = Probability Player A beats Player B

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \beta_2 X_2$$

$X_1 = \text{Rank}_{\text{Player A}} - \text{Rank}_{\text{Player B}}$

$X_2 = \text{Points}_{\text{Player A}} - \text{Points}_{\text{Player B}}$

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Example: Women's Tennis

Fit using glm in R: Maximum likelihood estimation via Fisher's scoring

> logit <- glm(outcome ~ $X_1 + X_2$, data = WTAfit, family = "binomial")

> summary(logit)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.967   -1.101    0.000    1.101    2.967

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.647e-15  1.031e-02    0.00        1
RankDiff    -3.096e-03  1.254e-04  -24.68   <2e-16 ***
PtsDiff      2.516e-04  7.018e-06   35.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use to calculate 95% confidence intervals

β_0 → (Intercept)

β_1 → RankDiff

β_2 → PtsDiff

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Example: Women's Tennis

Predict using predict in R:

```
> predictions <- predict(logit, newdata = matches, type = "link", se = TRUE)
> prob <- plogis(fit)
> LL <- plogis(fit – (1.96 * se.fit))
> UL <- plogis(fit + (1.96 * se.fit))
```

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Example: Women's Tennis

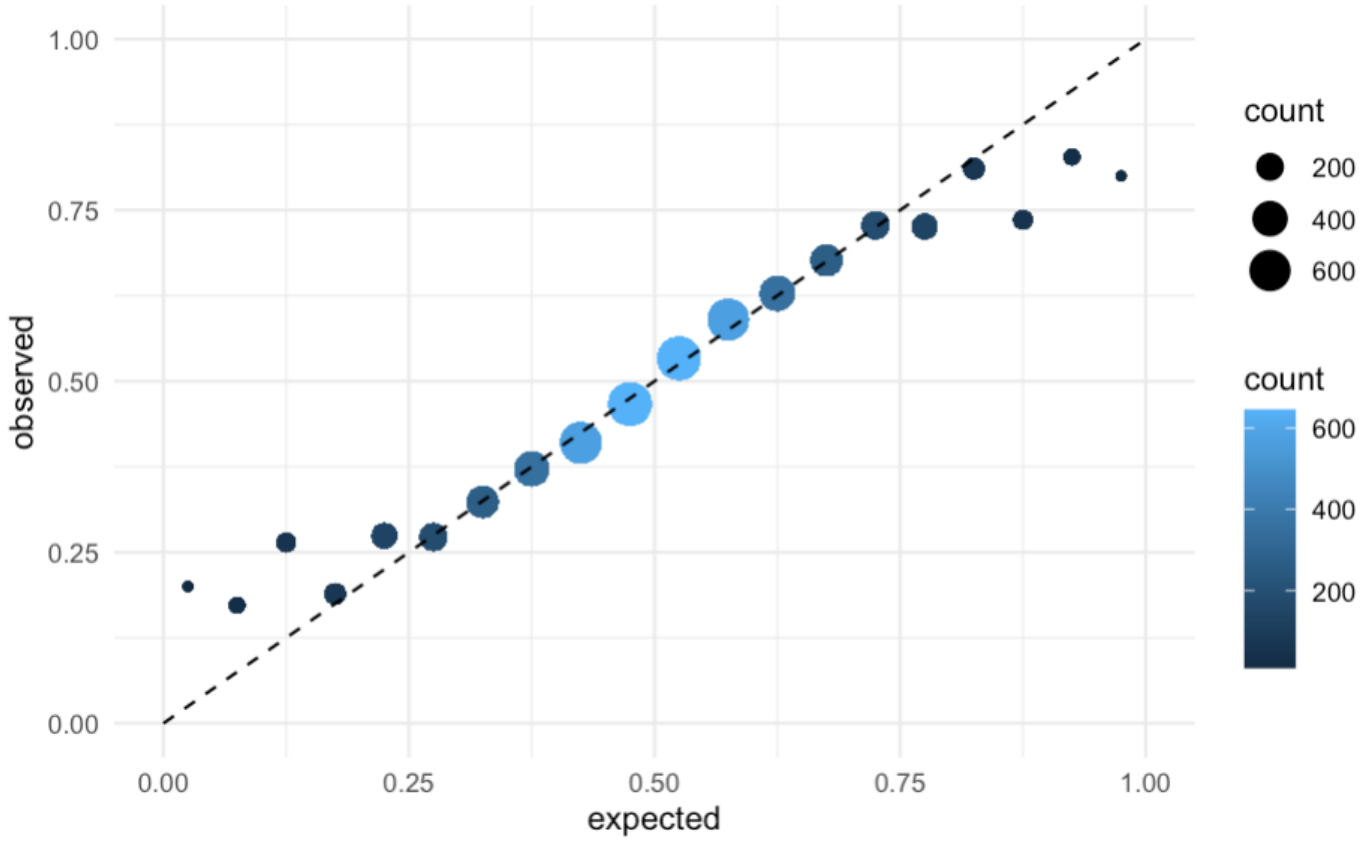
How well does the model fit?

Using the testing data

- Use the model to estimate win probabilities for all matches
- Bin by predicted win probability

  "expected"
- Plot against proportion matches won

  "observed"



Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Example: Women's Tennis

So, let's do some predictions!

What about 2023 Grand Slam Finals?

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg

Rybakina vs. Sabalenka
Australian Open 2023

Rybakina vs. Sabalenka

Independent variables: Rank diff $= 20,$ Pts diff $= -2755$

Note: **Odds** presented here are **decimal betting odds** $= \frac{1}{\pi}$

Rybakina vs. Sabalenka

Independent variables: Rank diff $= 20$, Pts diff $= -2755$

$\mathbb{P}[\text{Rybakina wins}] = 0.320\ (0.311, 0.329)$

Estimated odds $= 3.13\ (3.04, 3.22)$

Best available odds $= 2.22$

$\mathbb{P}[\text{Sabalenka wins}] = 0.680\ (0.671, 0.689)$

Estimated odds $= 1.47\ (1.45, 1.49)$

Best available odds $= 1.86$

Winner: Aryna Sabalenka

$\mathbb{P}$[Rybakina wins] $= 0.320\ (0.311, 0.329)$

Estimated odds $= 3.13\ (3.04, 3.22)$

Best available odds $= 2.22$

$\mathbb{P}$[Sabalenka wins] $= 0.680\ (0.671, 0.689)$

Estimated odds $= 1.47\ (1.45, 1.49)$

Best available odds $= 1.86$


Rank: 1st
Pts: 8940
Rank: 43rd
Pts: 1125
Swiatek vs. Muchova
French Open 2023

Swiatek vs. Muchova

Independent variables: Rank diff $= -42$, Pts diff $= 7815$

Swiatek vs. Muchova

Independent variables: Rank diff $= -42$, Pts diff $= 7815$

$\mathbb{P}[\text{Swiatek wins}] = 0.891\ (0.880, 0.900)$

Estimated odds $= 1.12\ (1.11, 1.14)$

Best available odds $= 1.19$

$\mathbb{P}[\text{Muchova wins}] = 0.109\ (0.100, 0.120)$

Estimated odds $= 9.14\ (8.33, 10.0)$

Best available odds $= 7.00$

Winner: Iga Swiatek

$\mathbb{P}[\text{Swiatek wins}] = 0.891\ (0.880, 0.900)$

Estimated odds $= 1.12\ (1.11, 1.14)$

Best available odds $= 1.19$

$\mathbb{P}[\text{Muchova wins}] = 0.109\ (0.100, 0.120)$

Estimated odds $= 9.14\ (8.33, 10.0)$

Best available odds $= 7.00$

Jabeur vs. Vondrousova
Wimbledon 2023

Jabeur vs. Vondrousova

Independent variables: Rank diff $= -36$, Pts diff $= 2351$

Jabeur vs. Vondrousova

Independent variables: Rank diff $= -36$, Pts diff $= 2351$

$\mathbb{P}$[Jabeur wins] $= 0.669\ (0.661, 0.677)$

Estimated odds $= 1.50\ (1.48\ 1.51)$

Best available odds $= 1.50$

$\mathbb{P}$[Vondrousova wins] $= 0.331\ (0.323, 0.339)$

Estimated odds $= 3.02\ (2.95, 3.09)$

Best available odds $= 3.02$

Winner: Marketa Vondrousova

$\mathbb{P}[\text{Jabeur wins}] = 0.669\ (0.661, 0.677)$

Estimated odds $= 1.50\ (1.48\ 1.51)$

Best available odds $= 1.50$

$\mathbb{P}[\text{Vondrousova wins}] = 0.331\ (0.323, 0.339)$

Estimated odds $= 3.02\ (2.95, 3.09)$

Best available odds $= 3.02$

Sabalenka vs. Gauff
US Open 2023

Sabelenka vs. Gauff

Independent variables: Rank diff $= -4$, Pts diff $= 4151$

Sabelenka vs. Gauff

Independent variables: Rank diff $= -4$, Pts diff $= 4151$

$\mathbb{P}$[Sabalenka wins] $= 0.742\ (0.730, 0.753)$

Estimated odds $= 1.35\ (1.33\ 1.37)$

Best available odds $= 1.83$

$\mathbb{P}$[Gauff wins] $= 0.258\ (0.247, 0.270)$

Estimated odds $= 3.88\ (3.71, 4.06)$

Best available odds $= 2.19$

Winner: Coco Gauff

$\mathbb{P}$[Sabalenka wins] $= 0.742\ (0.730, 0.753)$

Estimated odds $= 1.35\ (1.33\ 1.37)$

Best available odds $= 1.83$

$\mathbb{P}$[Gauff wins] $= 0.258\ (0.247, 0.270)$

Estimated odds $= 3.88\ (3.71, 4.06)$

Best available odds $= 2.19$

In summary

- Logistic regression can be used for predicting random variables with **binary** outcomes
- We can use the logistic function to link **probability of success** (or expected outcome) and **linear** combinations of **predictors**
- **Odds** describe the ratio between probabilities of success and failure
- We then fit a model using linear regression between the log odds and predictors
- We can apply these methods, e.g. in R

Code tutorial available at: github.com/emmalouisedavis/WTAlogisticreg