Modeling Task $f = f(I, R)$

* Statistical Models vs. Mathematical Models
(Machine Learning)

- Statistical/Machine learning Models are fit from historical data
- Mathematical Models are equations written on paper

Models fit from historical data

- What data do we need? Can we get it? get it.
- Choose the model

Mathematical Models

- What Random variables / distributions

are appropriate, if any?

- Statistical/Machine learning Models are fit from historical data

Models fit from historical data

- What data do we need? Can we get it?
- Choose the model

$f = f(I, R) =$ starting pit allows R runs through I innings, win prob.

data

every starting-pitcher-game → index $i = 1, \ldots, n$

R_i = runs allowed by s.p. in game i

I_i = innings pitched by s.p. in game i

(for simplicity assume s.p. pulled after fully completing I_i)

$y_i = 1$ if s.p. team wins, else 0

Lahman → box score

Retrosheet/Statcast → PBP

empirical grid

$$\hat{f}(I,R) = \text{Mean}\{y_i : I = I_i \text{ and } R = R_i\}$$

(I, R)

$(4, 2)$ 70% win, 30% loss $\approx .7$

XGBoost with Monotonic Constraints

XGBoost = one of the fastest and easiest to use "off-the-shelf" supervised machine learning algorithms

dependent variables $X_1, \ldots, X_p$
independent variables y

XGBoost "memorizes" as best as possible

$$g(X) = y$$

interpolates

Monotonic constraints:

$R \rightarrow$ tell XGBoost to be monotonic decreasing in R

$I \rightarrow$ tell XGBoost to be monotonic increasing in I

Try Mathematical Models

$(R, I) \rightarrow$ win probability

say inning i

X_i = # runs scored by the pitcher's team

Y_i = # runs scored by the opp's team

$$\hat{f}(I, R) = \mathbb{P}\left(\sum_{i=1}^{9} X_i > R + \sum_{i=I+1}^{9} Y_i\right) + \frac{1}{2}\,\mathbb{P}\left(\sum_{i=1}^{9} X_i = R + \sum_{i=I+1}^{9} Y_i\right)$$

$\boxed{\text{Start simple}} \rightarrow$ Poisson

X_i, Y_i random variables (distributions)

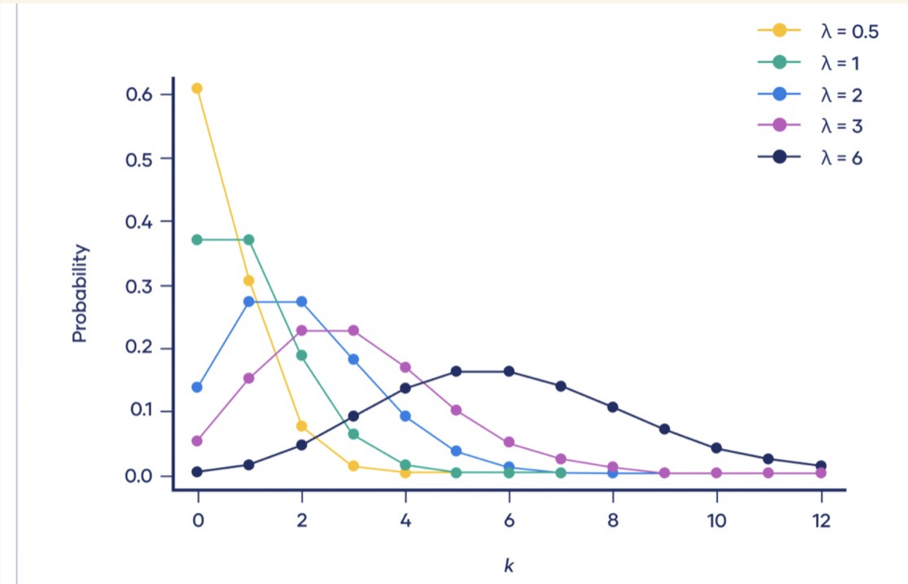
↳ possible outcomes $0, 1, 2, 3, \ldots$

$X \sim \text{Poisson}(\lambda)$ means

outcomes $x \in \{0,1,2,3,\dots\}$

$$\mathbb{P}(X = x) = e^{-\lambda x} \frac{\lambda^{x}}{x!}$$

$\mathbb{E}X = \lambda$, $\text{var}(X) = \lambda$, $\lambda > 0$



$$\mathbb{P}\left(\sum_{i=1}^{9} X_i > R + \sum_{i=I+1}^{9} Y_i \right)$$

$$= \left\{ \mathbb{P}\left(\sum_{i=1}^{9} \text{Poisson}(\lambda) > R + \sum_{i=I+1}^{9} \text{Poisson}(\lambda) \right) \quad \text{if } I < 9 \right.$$

$$\left(\mathbb{P}\left(\sum_{i=1}^{9} \text{Poisson}(\lambda) > R \right) \quad \text{if } I = 9 \right.$$

$$= \begin{cases} \mathbb{P}\left[\text{Poisson}(9 \cdot \lambda) > R + \text{Poisson}\big((9-(I+1))\lambda\big) \right] \\ \mathbb{P}\left(\text{Poisson}(9 \cdot \lambda) > R \right) & \text{if } I = 9 \end{cases}$$

<u>Thm</u> $\text{Poisson}(\lambda_1) + \text{Poisson}(\lambda_2) = \text{Poisson}(\lambda_1 + \lambda_2)$

$$= \begin{cases} \mathbb{P}\left[\text{Skellam}(9\lambda, (9-I-1)\lambda) > R \right] & \text{if } I < 9 \\ \mathbb{P}\left[\text{Poisson}(9\lambda) > R \right] & \text{if } I = 9 \end{cases}$$

<u>Thm</u> $\text{Poisson}(\lambda_1) - \text{Poisson}(\lambda_2)$
$= \text{Skellam}(\lambda_1, \lambda_2)$

Now have an explicit formula for $f(I,R)$ if we assume

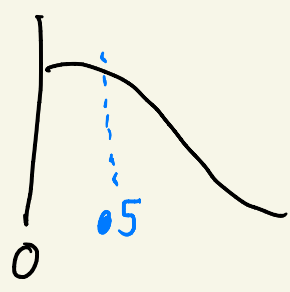
$$\boxed{X_i, Y_i \sim \text{Poisson}(\lambda).}$$

$\hat{\lambda} = \mathbb{E}X_i =$ mean runs allowed in an inning by one team

Need to choose a smart value of λ

for a given league-season (e.g. 2019 NL)
let $\lambda =$ observed mean runs allowed in a half-inning

Code



$$X_i \sim \text{Poisson}(\lambda_x)$$

$$Y_i \sim \text{Poisson}(\lambda_y)$$

$$\lambda_x^{(m)}, \lambda_y^{(m)} \sim \mathcal{N}_+(\lambda, \sigma^2)$$

$$f(R, I \mid \lambda_x, \lambda_y) =$$

$$\mathbb{P}\left(\sum_{i=1}^{9} X_i > R + \sum_{i=I+1}^{9} Y_i \right)$$

$$+ \frac{1}{2} \mathbb{P}\left(\sum_{i=1}^{9} X_i = R + \sum_{i=I+1}^{9} Y_i \right)$$

= a similar formula as before except now in terms of λ_x, λ_y

$$f(I, R) = \frac{1}{M} \sum_{m=1}^{M} f(I, R \mid \lambda_x^{(m)}, \lambda_y^{(m)})$$

$$\begin{cases} X_i \sim \text{Poisson}(\lambda_x) \\ Y_i \sim \text{Poisson}(\lambda_y) \\ \lambda_x^{(m)}, \lambda_y^{(m)} \sim N_+(\lambda, \sigma^2) \end{cases}$$

$$\lambda = \text{mean}\left(\left\{ \begin{matrix} \text{mean Runs scored in} \\ \text{half inning by} \\ \text{team } t \end{matrix} \right\} \right)$$

$$\sigma^2 = \text{var}\left(\left\{ \; ? \; \right\} \right)$$

2019 NL: get $\hat{\lambda}, \hat{\sigma}^2$

then

$$f(I, R \mid \hat{\lambda}, \hat{\sigma}^2) = \frac{1}{M} \sum_{m=1}^{M} f(I, R \mid \lambda_x^{(m)}, \lambda_y^{(m)})$$

where $\lambda_x^{(m)}, \lambda_y^{(m)} \sim N_+(\hat{\lambda}, \hat{\sigma}^2)$

$M = 100$

$$\begin{cases} X_i \sim \text{Poisson}(\lambda_x) \\ Y_i \sim \text{Poisson}(\lambda_y) \\ \lambda_x^{(m)}, \lambda_y^{(m)} \sim N_+(\lambda, \sigma^2 \cdot K) \end{cases}$$

$K < 1$

$$f(I, R \mid \hat{\lambda}, \hat{\sigma}^2, K) = \frac{1}{M} \sum_{m=1}^{M} f(I, R \mid \lambda_x^{(m)}, \lambda_y^{(m)})$$

where $\lambda_x^{(m)}, \lambda_y^{(m)} \sim N_+(\hat{\lambda}, \hat{\sigma}^2 \cdot K)$

↳ grid $f(I, R \mid \hat{\lambda}, \hat{\sigma}^2, K)$

how to choose K?

try different K's and choose the K which ~~works~~ works best, meaning makes the most accurate predictions.

WP predictions $f(I, R \mid \hat{\lambda}, \hat{\sigma}^2, K)$

observed Win/loss column

logloss$\cdot$(pred, obs. win/loss)