Nearest-Neighbor Matchup Effects: Predicting March Madness

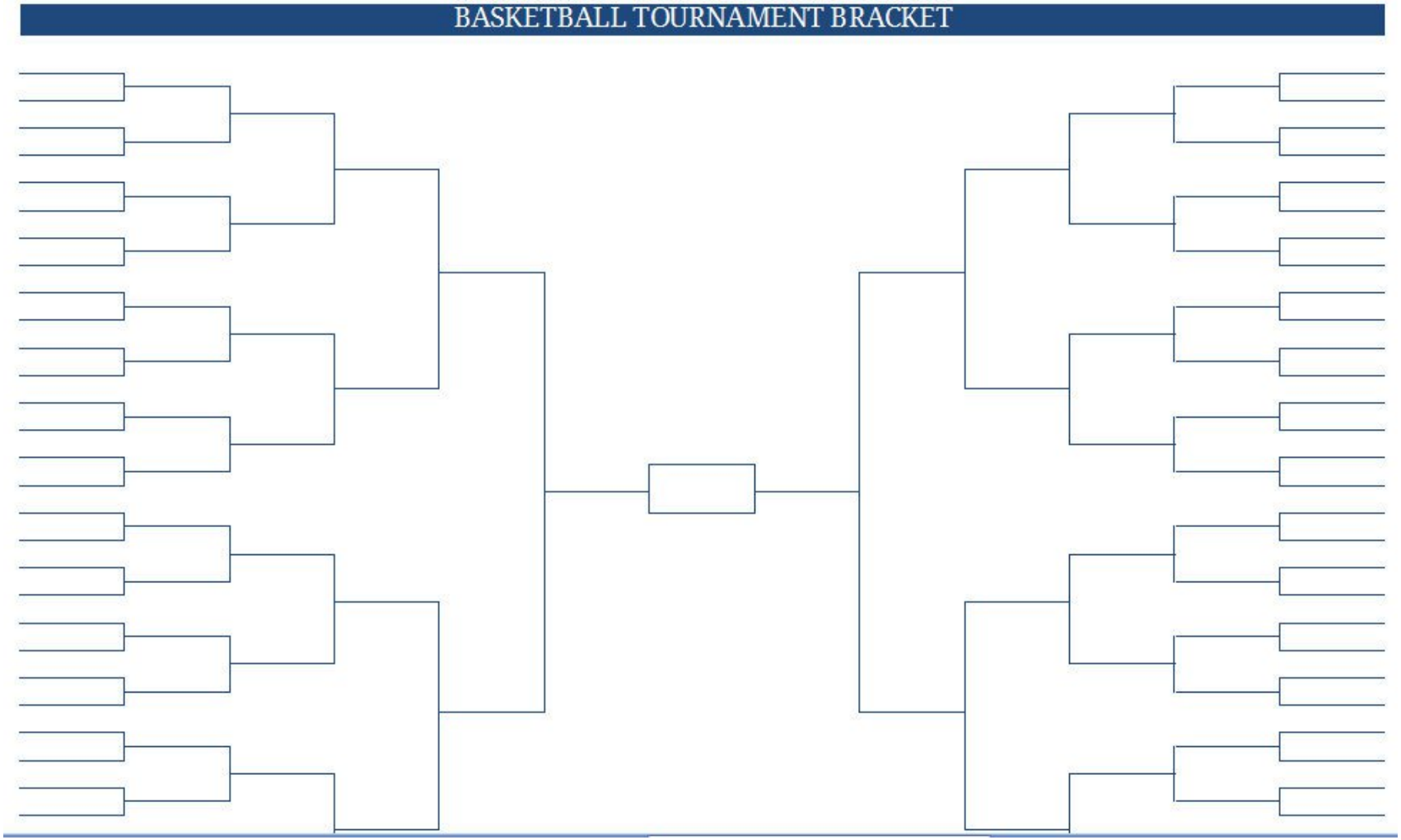
Andrew Hoegh, Marcos Carzolio, Ian Crandell, Xinran Hu,
Lucas Roberts, Yuhyun Song, Scotland Leman

Department of Statistics, Virginia Tech

September 26, 2015

NCAA Bracket Competition

Who is has filled out an NCAA bracket before?



NCAA Bracket Competition

Who has won an NCAA bracket competition?

NCAA Bracket Competition

Who has lost a bracket competition to someone that does not know what a basketball looks like?

Talk Overview

- General modeling strategies for NCAA tournament competitions
- Matchup effects modeling framework
- Uncertainty inherent in NCAA tournament & competitions

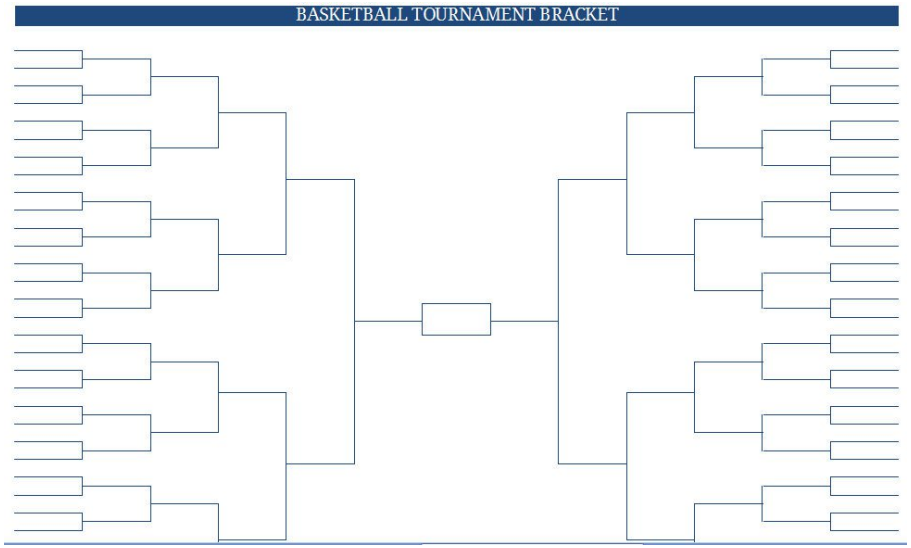
Competition Types

kaggle™

Loss Functions: Bracket

$$\mathcal{L}(y, \hat{y}) = c\delta(y \neq \hat{y})$$

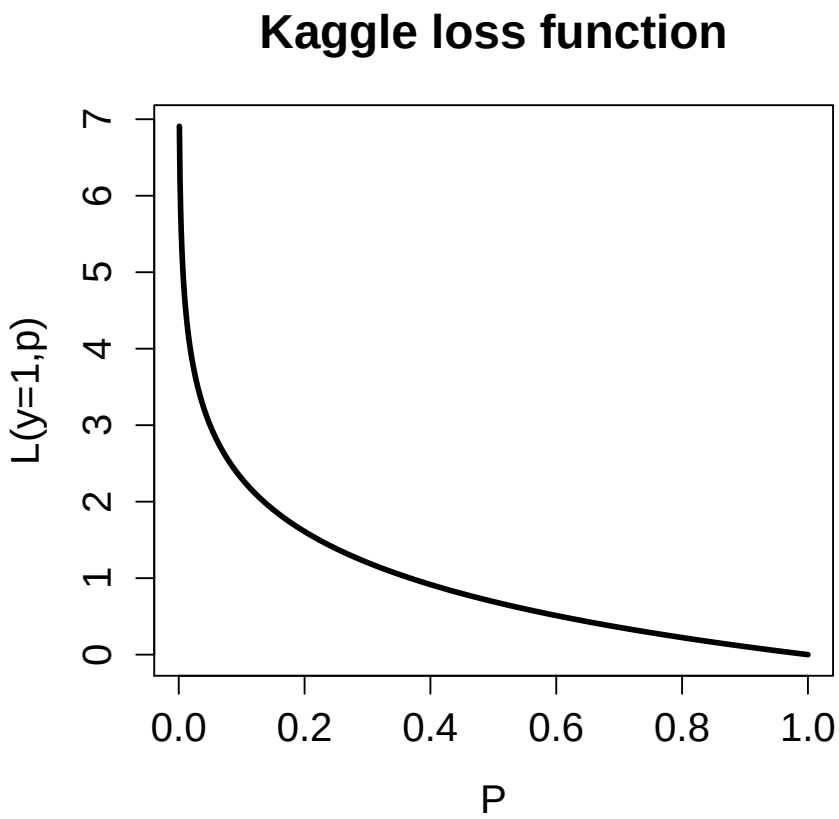
$$\mathcal{L}(y, \hat{y}, r) = c_r\delta(y \neq \hat{y})$$



Loss Functions: Probabilistic Prediction

$$\mathcal{L}(y, p) = -y \log(p) - (1-y) \log(1-p)$$

This is a *proper* scoring rule.



Relevant Data: Team Characteristics

There are many important characteristics useful for modeling the strength of an NCAA basketball team:

- Winning percentage,
- Point differential,
- Strength of schedule,
- Conference affiliation,
- ...
- Rebounding percentage,
- Adjusted offensive efficiency, and
- Adjusted defensive efficiency.

Relevant Data: Ratings and Rankings

Rather than using team characteristics, there are many available rating and ranking systems:

- ESPN BPI,
- Sagarin,
- RPI,
- Pomeroy, and
- Logistic Regression/Markov Chain (LRMC).

Relative Strength Models

Typically the model framework for predicting winner of games (or winning probabilities) can be formulated as a relative strength model. Formally this can be expressed as:

$$y_{ij} \quad = \quad f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$$

Relative Strength Models

Typically the model framework for predicting winner of games (or winning probabilities) can be formulated as a relative strength model. Formally this can be expressed as:

$$\begin{aligned} y_{ij} &= f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \\ &\text{such as} \\ \text{linear model: } y_{ij} &= \beta_{home} + (\theta_i - \theta_j)\beta_D + \epsilon_{ij}, \\ \text{where} & \quad \epsilon_{ij} \sim N(0, \sigma^2) \end{aligned}$$

Relative Strength Models

Typically the model framework for predicting winner of games (or winning probabilities) can be formulated as a relative strength model. Formally this can be expressed as:

$$
\begin{aligned}
y_{ij} &= f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \\
&\text{such as} \\
\text{linear model: } y_{ij} &= \beta_{home} + (\theta_i - \theta_j)\beta_D + \epsilon_{ij}, \\
\text{where} \quad & \epsilon_{ij} \sim N(0, \sigma^2) \\
&\text{or} \\
\text{logistic regression: } y_{ij} &\sim Bernoulli(p_{ij}) \\
\text{where} \quad & logit(p_{ij}) = \beta_{home} + (\theta_i - \theta_j)\beta_D
\end{aligned}
$$

Transitivity

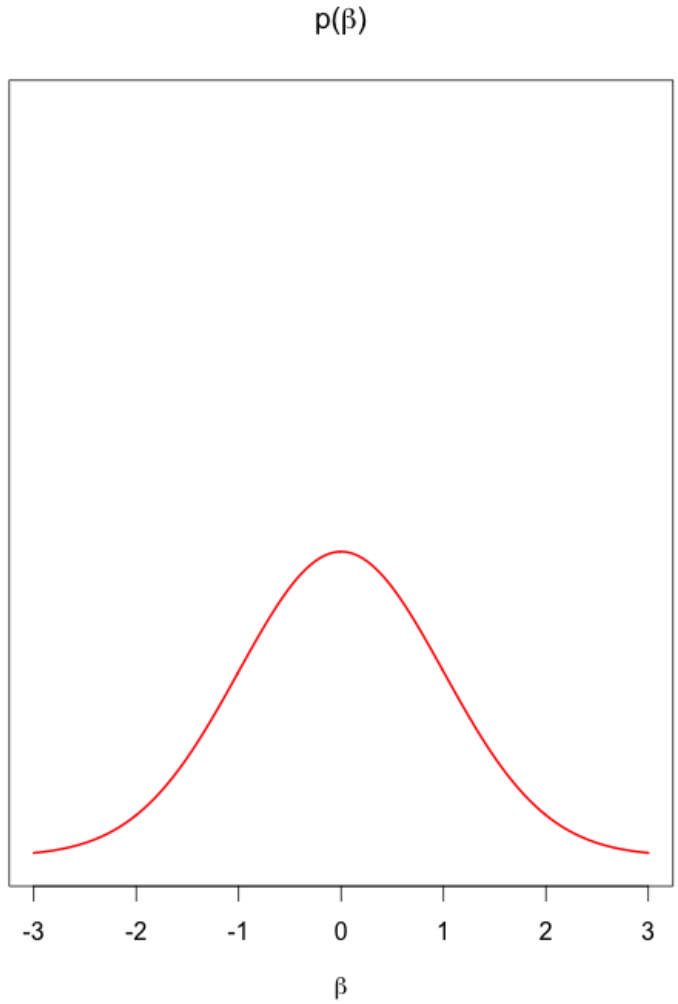
Note that relative strength models of this type are strictly transitive, where $P_{A>B}$ denotes the probability that team A beats team B. Under transitive models,

$$\{P_{A>B} > 0.5 \cup P_{B>C} > 0.5\} \Rightarrow P_{A>C} > 0.5.$$

Bayesian Relative Strength Models

Bayesian Linear Model

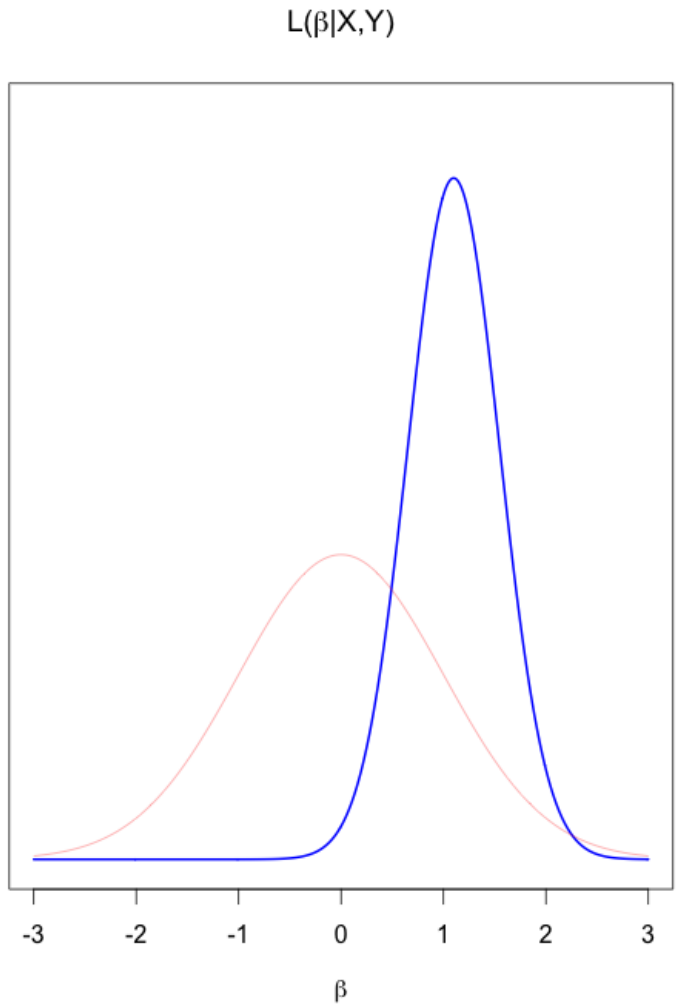
Prior	$\beta \sim N(\beta; m, s)$
Likelihood	$\beta \vert\, Y, X \propto N(\beta; \hat{\beta}, \Sigma)$
Posterior	$\beta \vert\, Y, X \sim N(\beta; \mu, \Sigma_\beta)$



Bayesian Relative Strength Models

Bayesian Linear Model

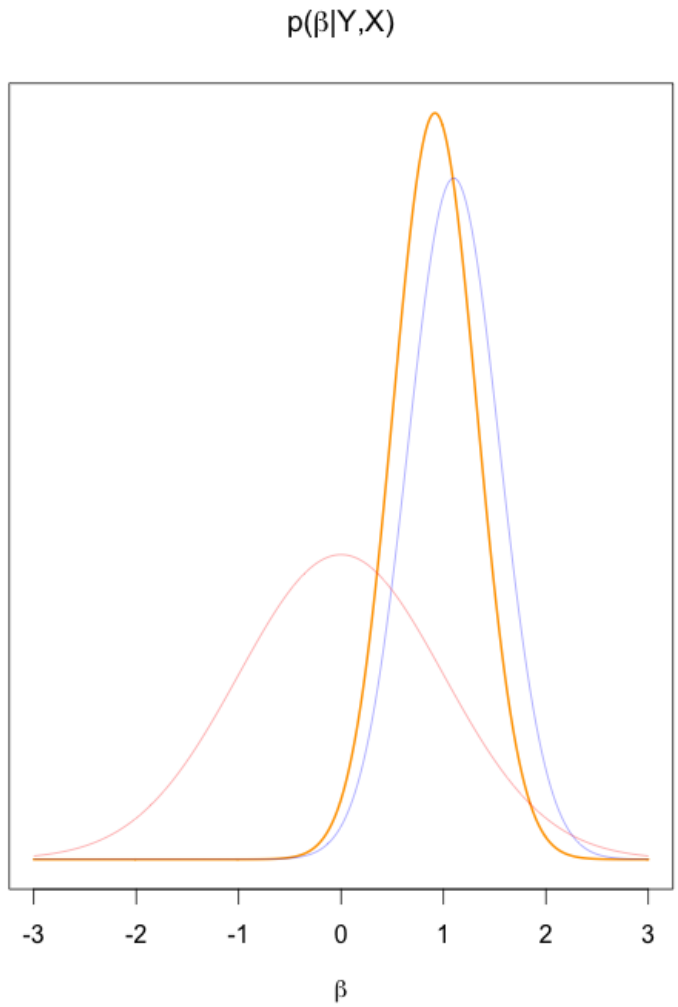
Prior	$\beta \sim N(\beta; m, s)$
Likelihood	$\beta \mid Y, X \propto N(\beta; \hat{\beta}, \Sigma)$
Posterior	$\beta \mid Y, X \sim N(\beta; \mu, \Sigma_\beta)$



Bayesian Relative Strength Models

Bayesian Linear Model

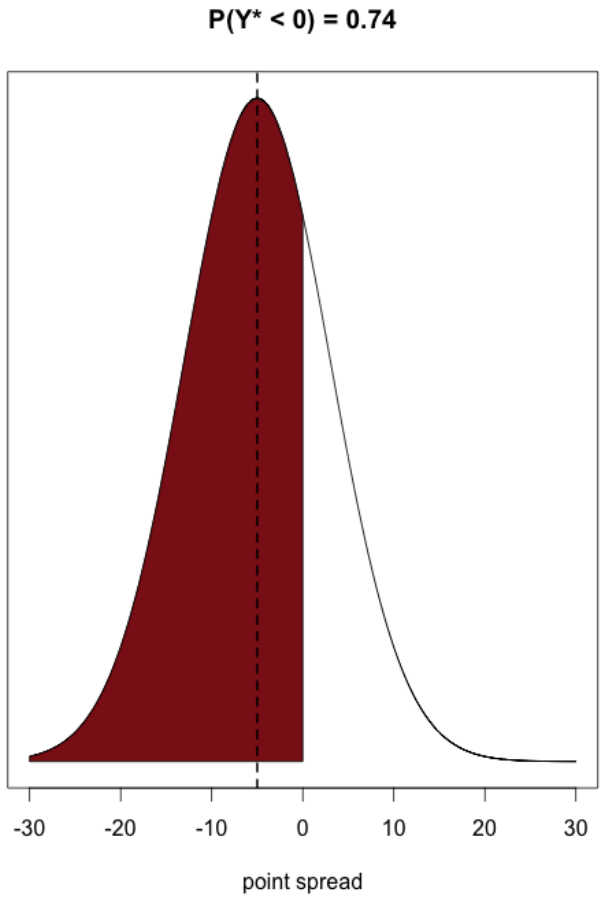
Prior	$\beta \sim N(\beta; m, s)$
Likelihood	$\beta \vert\, Y, X \propto N(\beta; \hat{\beta}, \Sigma)$
Posterior	$\beta \vert\, Y, X \sim N(\beta; \mu, \Sigma_\beta)$



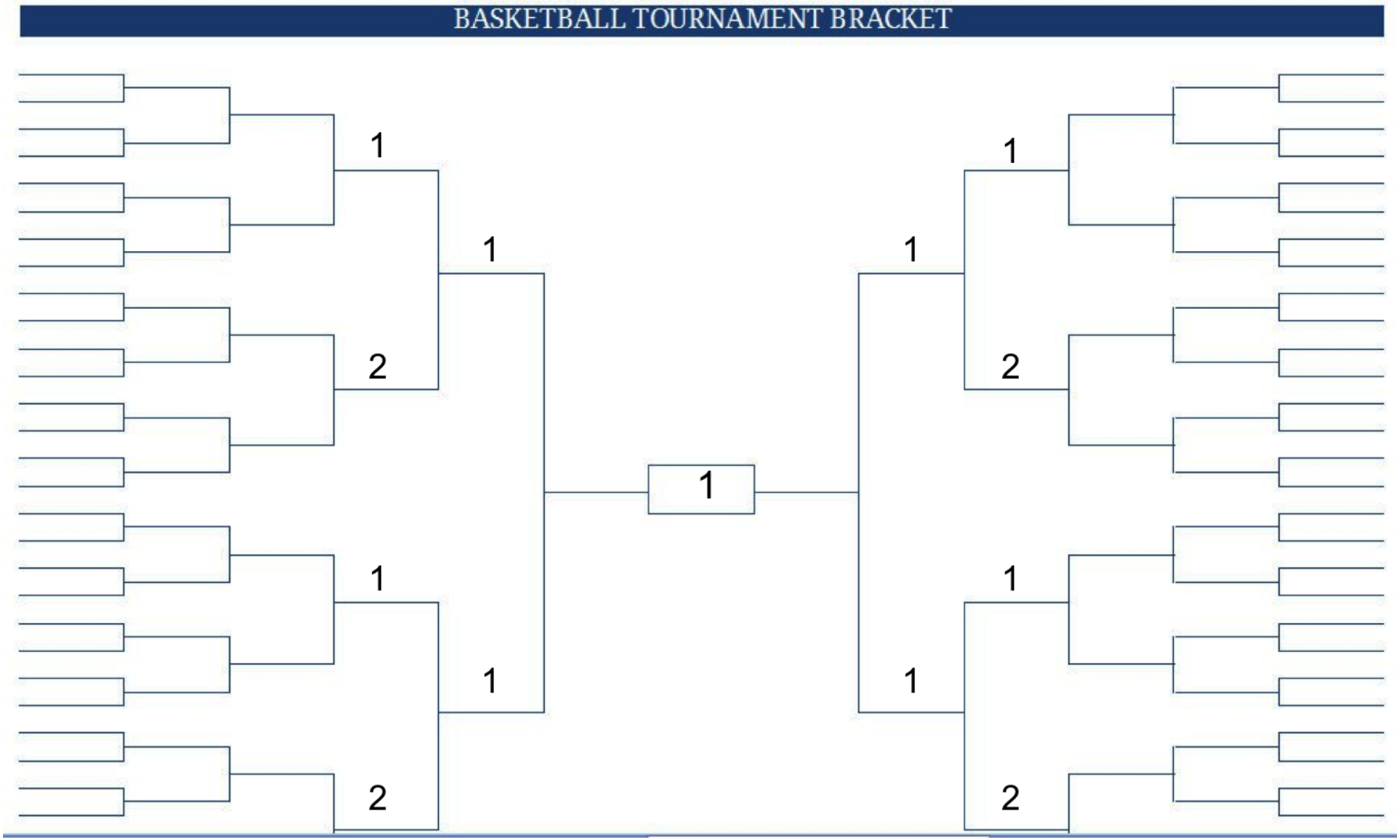
Bayesian Relative Strength Models

Bayesian Linear Model

$$\begin{aligned} \text{Predictive:} \quad & p(Y^*|X^*, Y, X) \\ = \quad & \int p(Y^*|X^*, \beta) p(\beta|Y, X) d\beta \\ P(Y^* < 0) \quad = \quad & \int_{-\infty}^{0} p(Y^*|X^*, Y, X) dY^* \end{aligned}$$



Analytics for Bracket Competitions

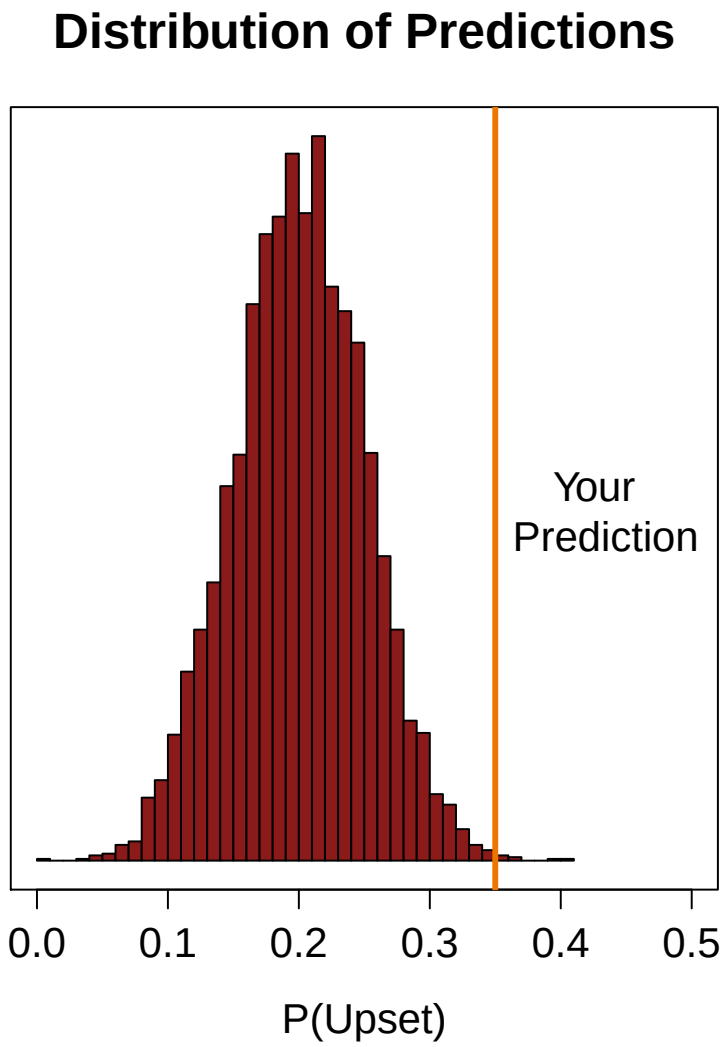


Analytics for Kaggle-Style Competitions

A substantial challenge in predictive modeling of sports competitions, and major discussion point, is predicting upsets. Consider "predicting upsets" as a comparison of the probabilistic predictions for competitors.

Analytics for Kaggle-Style Competitions

A substantial challenge in predictive modeling of sports competitions, and major discussion point, is predicting upsets. Consider "predicting upsets" as a comparison of the probabilistic predictions for competitors.

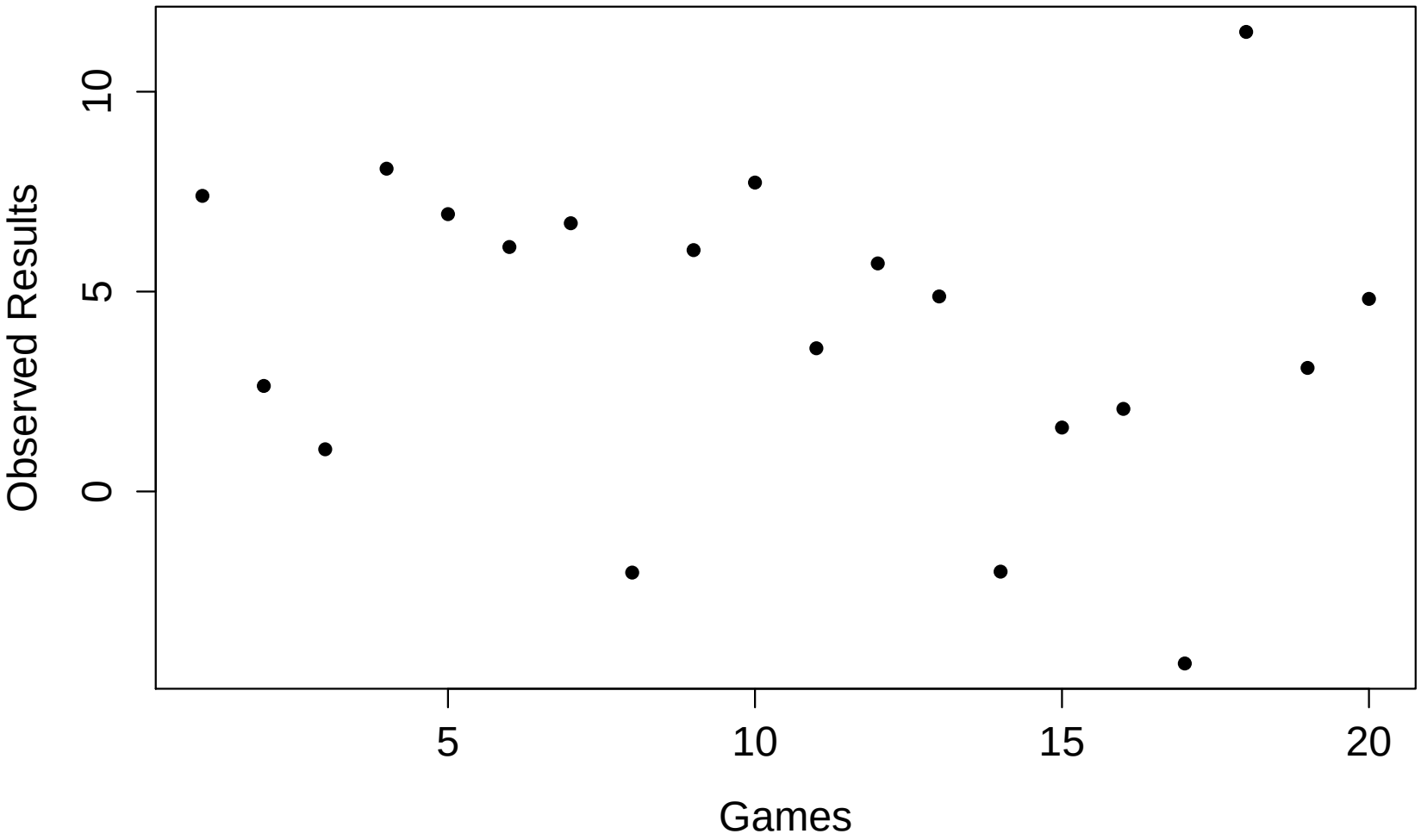


Matchup Effects Motivating Quote

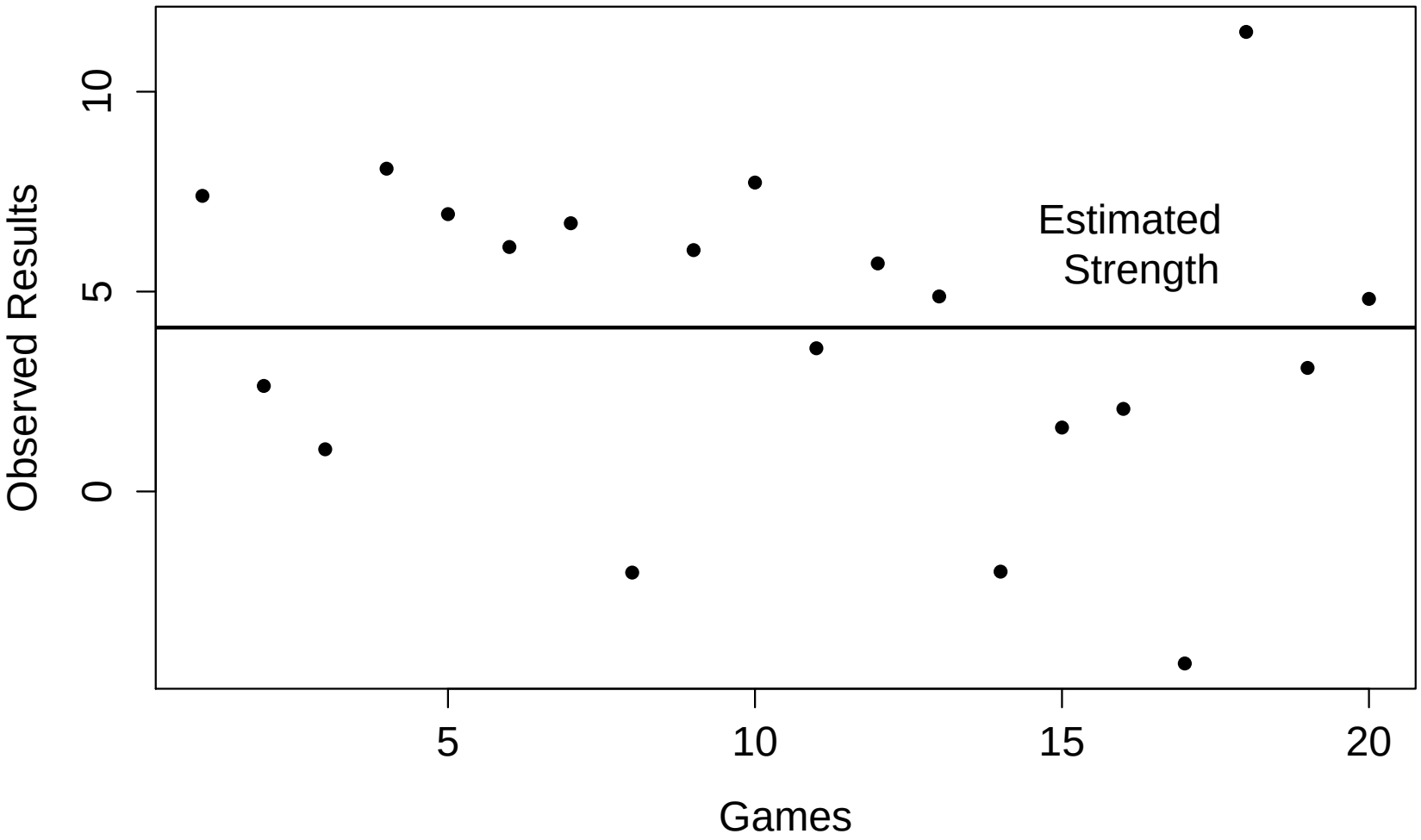
"Well, it's hard to predict a particular team. You have to look at the higher seed and see, do they overwhelm the smaller school or the lower seed? Can they overwhelm someone with their athleticism or length or size or quickness or speed? Do they play a particular style of the pressure defense? And I think I look at the lower seed then and say, can they counter the higher seed's strengths? Do they shoot the three point shot well? Do they have athleticism at certain positions? Do they play a particular style that will give the higher seed trouble?"

Andy Enfield

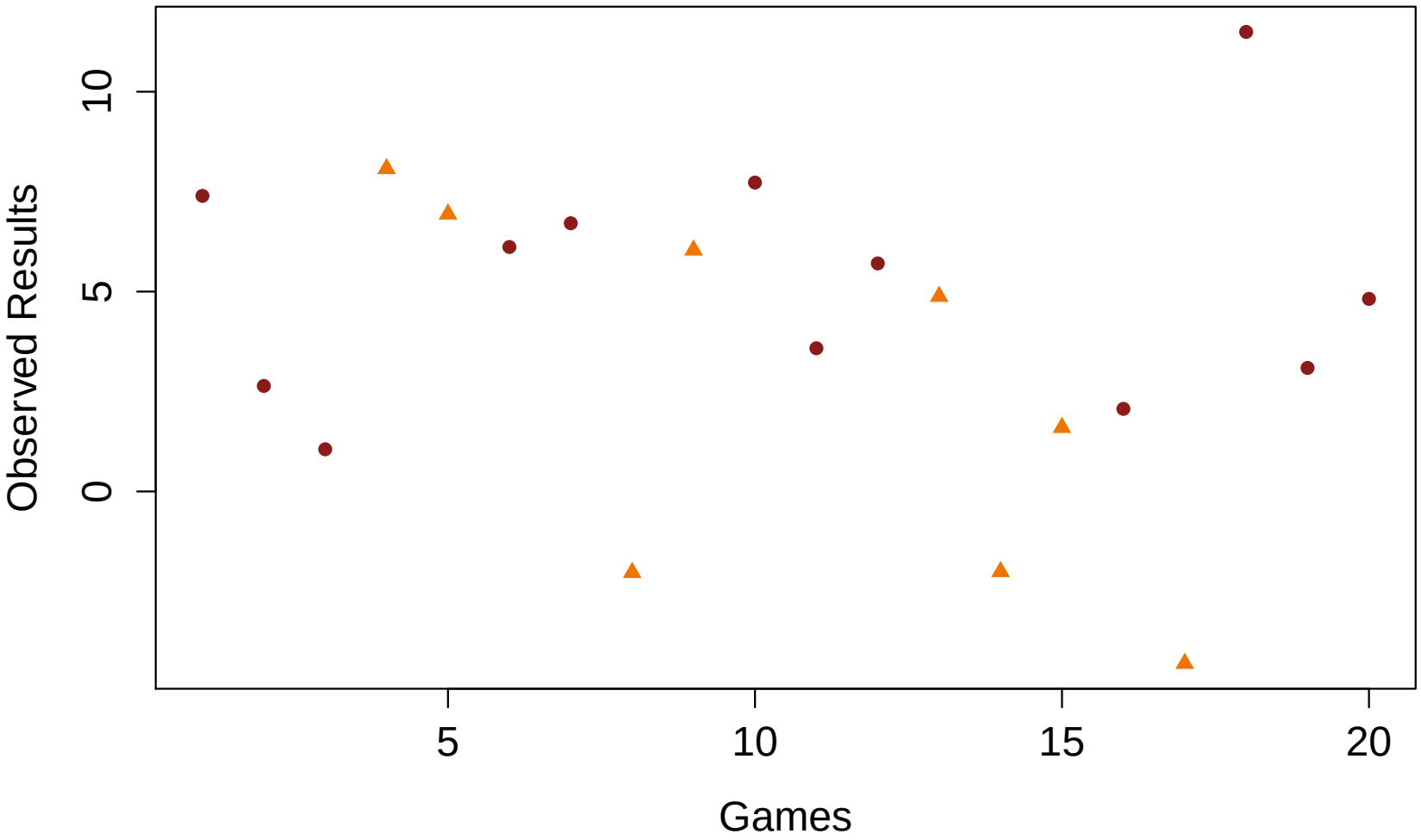
NNME Intuition



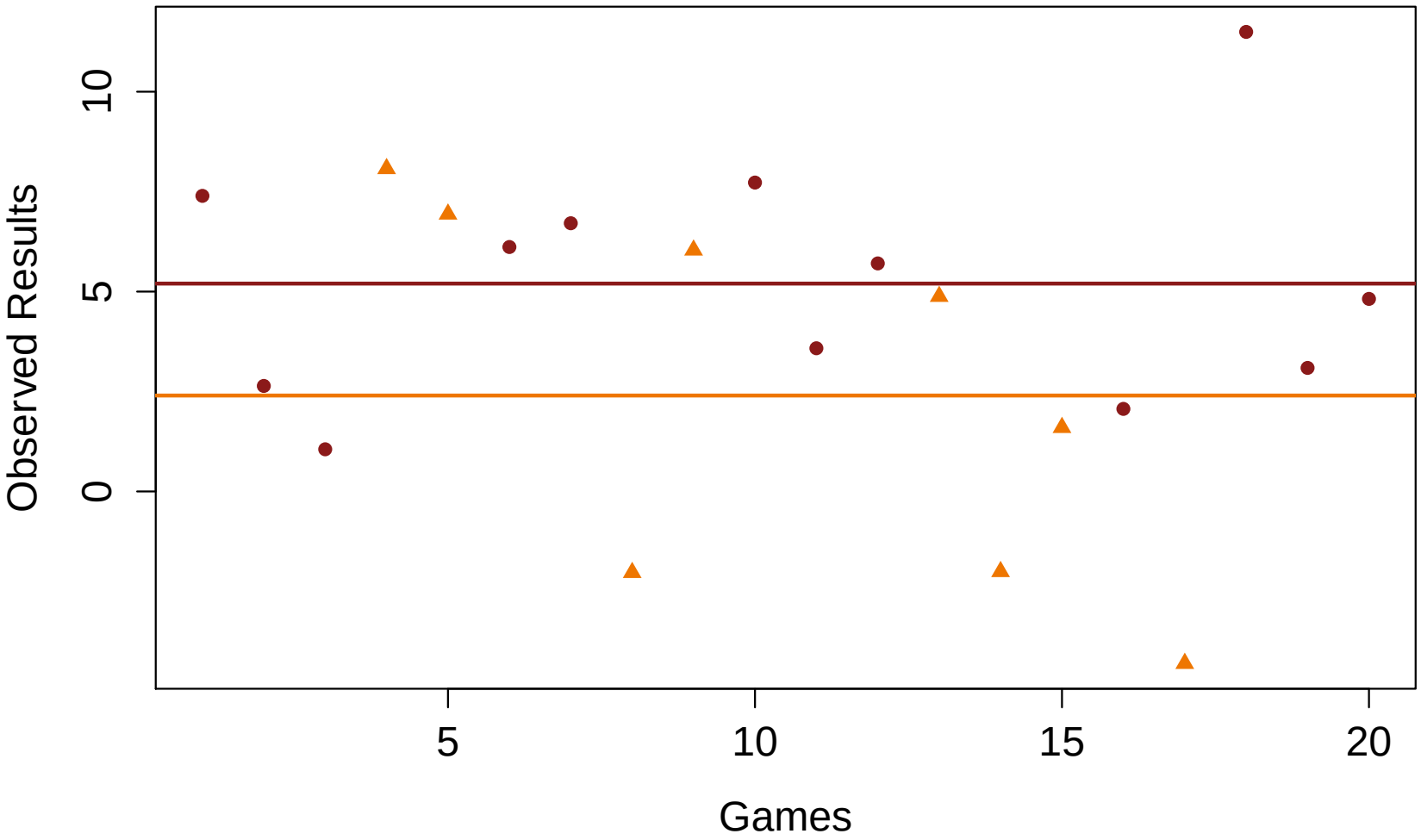
NNME Intuition



NNME Intuition



NNME Intuition



Outline of the NNME

Estimation of the nearest-neighbor matchup effects has three components:

1. Fit a relative strength model,
2. Identify neighbors for the matchup, and
3. Calibrate the matchup adjustment.

Relative Strength Model

Consider the simple relative strength model using the Sagarin ratings and home court effect.

$$\begin{aligned} Y_{ij} &= \beta_{home} + D_{ij}\beta_D + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2) \\ \text{where} \quad & D_{ij} = \text{Sagarin}_i - \text{Sagarin}_j \end{aligned}$$

Relative Strength Model

Consider the simple relative strength model using the Sagarin ratings and home court effect.

$$\begin{aligned} Y_{ij} &= \beta_{home} + D_{ij}\beta_D + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2) \\ \text{where} \quad & D_{ij} = \text{Sagarin}_i - \text{Sagarin}_j \end{aligned}$$

Note that relative strength models of this type are strictly transitive.

Identifying Neighbors

Identifying neighbors first requires specifying team characteristics and a distance function between teams. We used a collection of data from Ken Pomeroy's `www.kenpom.com` including:

- Effective Height
- Adjusted Tempo
- Effective Field Goal Percentage Defense
- Offensive Rebound Percentage
- Block Percentage
- Steal Rate
- Three Point Field Goal Contribution

Uniformly Weighting Characteristics

Bayesian Visual Analytics

Bayesian Visual Analytics

Finding Neighbors

For each particular matchup, say Dayton vs. Stanford, identify past opponents to see who is most like the current opponent.

Teams Dayton played similar to Stanford

- California
- George Mason
- Georgia Tech
- George Washington
- Gonzaga

Teams Stanford played similar to Dayton

- Cal Poly
- California
- Oregon
- Pittsburgh
- Utah

Matchup Adjustment Notation

Let $\mathcal{R}_j(i) = \frac{1}{K}\sum_{k \in \mathcal{N}_j}(Y_{ik} - \mu_{ik})$, where $\mathcal{N}_j$ are the neighbors for team i with respect to team j, Y_{ik} is the observed point differential between team i, and team k and μ_{ik} is the expected point differential between team i and team k.

Then $\phi_{ij} = \rho(\mathcal{R}_i(j) - \mathcal{R}_j(i))$, where ρ controls how much information is passed from the neighbors.

Predictive Distribution

Then using the relative strength model previously described, the predictive distribution becomes:

$$Y_{ij}|X_{ij} = \beta_{home} + D_{ij}\beta_D + \phi_{ij} + \epsilon_{ij}$$

Effectively ϕ_{ij} shifts the predictive distribution and results in a non-transitive model.

Estimating Model Parameters

Using data across NCAA tournaments from 2007 - 2013, the model parameters are estimated.

	β_{home}	β_D	σ^2	ρ
Posterior mean	3.87	0.913	121.6	0.167
Credible interval	(3.83,3.91)	(0.909,0.916)	(120.0,123.2)	(0.012,0.454)

The positive credible interval suggests a moderate, but meaningful result from the matchup effect.

Demonstration: 2014 NCAA Tournament

Largest shifts in expected point spread (ϕ_{ij}).

Team_1	Team_1	$\mathcal{R}_1(2)$	$\mathcal{R}_2(1)$	ϕ_{12}
Cal Poly	Wichita St.	7.52	-0.44	1.59
UConn	St. Joes	0.70	-8.80	1.90
Dayton	Stanford	14.97	-0.55	3.10
Dayton	Syracuse	8.65	-2.83	2.30
Kentucky	Michigan	2.2	-5.87	1.62
UMass	Tennessee	-1.75	7.16	-1.78
Memphis	Virginia	-8.09	4.74	-2.57
Michigan	Tennessee	-2.85	6.55	-1.88
Michigan	Texas	-5.87	5.12	-2.20
Syracuse	W.Mich	6.27	-5.52	2.36

Closer Look: 2014 NCAA Tournament

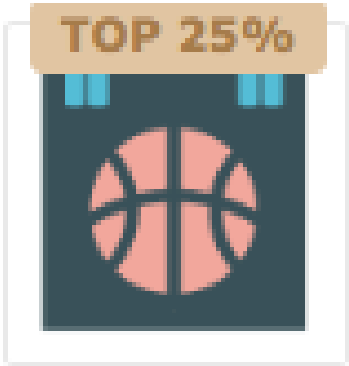
Team	Neighbor	Point Diff.	E[Point Diff]	Residual
Dayton	California	18	-0.9	18.9
Dayton	Gonzaga	5	-12.4	17.4
Dayton	George Mason	17	3.4	13.6
Dayton	Georgia Tech	10	-5.3	15.3
Dayton	George Washington	10	0.4	9.6
Stanford	California	-7	4.1	-11.1
Stanford	California	11	-2.3	13.3
Stanford	Oregon	2	-8.9	10.9
Stanford	Pittsburgh	-21	-5.3	-15.7
Stanford	Cal Poly	17	13.8	3.2
Stanford	Utah	1	4.9	-3.9

NNME Concluding Thoughts

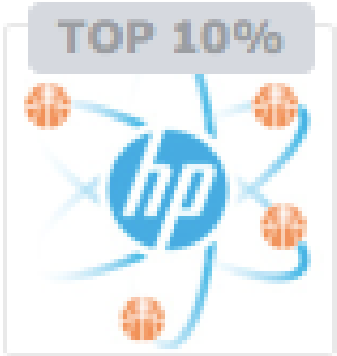
1. Evaluated on small number of data points leads to very uncertain outcomes, even with quality predictions
2. Does this contest actually have a proper scoring rule?
3. Alternative strategies (maximizing expected return).

NNME Concluding Thoughts

1. Evaluated on small number of data points leads to very uncertain outcomes, even with quality predictions
2. Does this contest actually have a proper scoring rule?
3. Alternative strategies (maximizing expected return).



59th/248



16th/341