Estimating the Probability of Winning a College Basketball Game

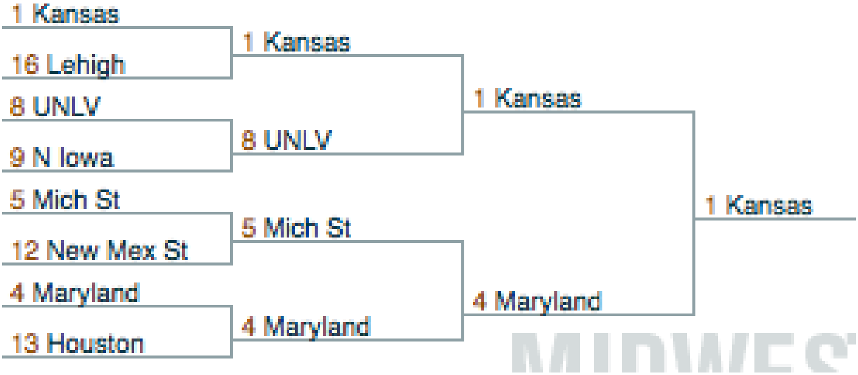
Ryan J. Parker
ryan@basketballgeek.com

College of Charleston

March 21, 2010

Why Win Probabilities Matter

- Classical ranking says the higher ranking team wins



Why Win Probabilities Matter: Answering Questions

- Who is the favorite to win the tournament?
- How often does Kentucky make the Final Four?
- Will St. Mary's advance to the Sweet 16?

Why Win Probabilities Matter: Taking Your Friend's $$

- Essential for non-standard point systems

Tournament Path Matters: Extreme Example

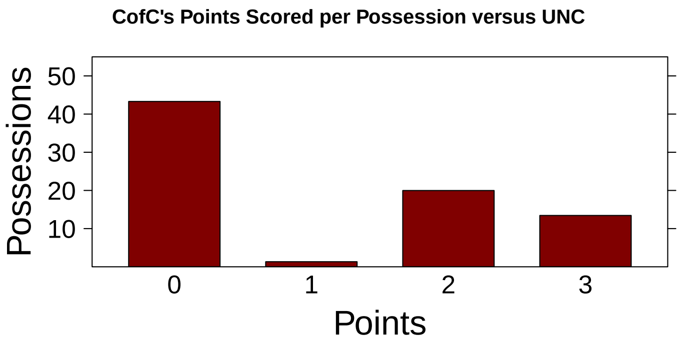
Rank		Win Probability
1	Team A	51%
2	Team B	49%
3	Team C	100%
4	Team D	0%

- Assume Team A and Team B beat Team C 65% of the time:
  - **Championship Odds**:
    Team A: 33%
    Team B: 32%
    Team C: 35%
- Classical ranking chooses Team A, but Team C has the best chance of winning

Actual Tournament Path Examples

- **2003**: #2 Kentucky over #1 Kansas to win championship
  - Kentucky 14%, Kansas 13%
- **2010**: #2 Kansas over #1 Duke to win championship
  - Kansas 24.8%, Duke 24.5%
- **2010**: #18 Villanova more likely than #15 Baylor to make Elite 8
  - Villanova 35%, Baylor 33%

Data to Rank: Points per Possession



- **Efficiency**: mean number of points scored per possession
- Removes effect of pace on a team's points scored and allowed
- Possessions estimated with FGA $-$ OR $+$ TO $+$ 0.475 $\times$ FTA

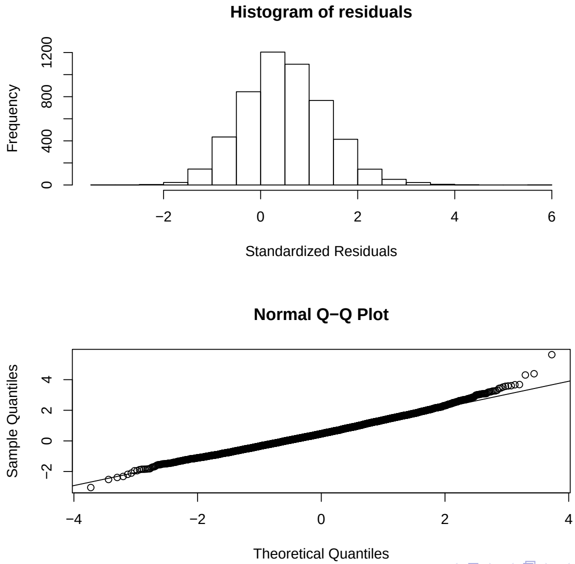
Linear Model of Difference in Per Game Efficiency

- D_{ij}: difference in team i's and team j's efficiency
- Linear regression where we assume

$$D_{ij} \sim \mathrm{N}(\alpha(\text{home}) + \beta_i - \beta_j, \sigma_d^2)$$

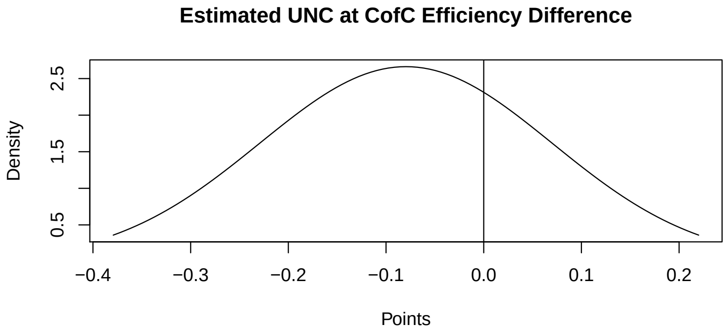
- β_i: rating for team i
- home$= \begin{cases} 1 & \text{if team } i \text{ is at home} \\ -1 & \text{if team } i \text{ is away} \\ 0 & \text{otherwise} \end{cases}$
- When $D_{ij} > 0$, team i beats team j

Assumption of Normality



Estimates from the Linear Model for 2010

- **Home Court**: $\hat{\alpha} = 0.05 = 3.5$ points
- **Standard Deviation**: $\hat{\sigma_d} = 0.15 = 10.5$ points



- Pr(CofC Win) = 0.30

Answering Questions with the Linear Model

- Who is the favorite to win the tournament?
  - Kansas, 25%
- How often does Kentucky make the Final Four?
  - 17%
- Will St. Mary's advance to the Sweet 16?
  - 25%

Multinomial Model of Point Probabilities

- Estimates probability of scoring points on possessions
- We consider 0, 1, 2, or $\geq$ 3 points
- **Multinomial Logistic Regression**:

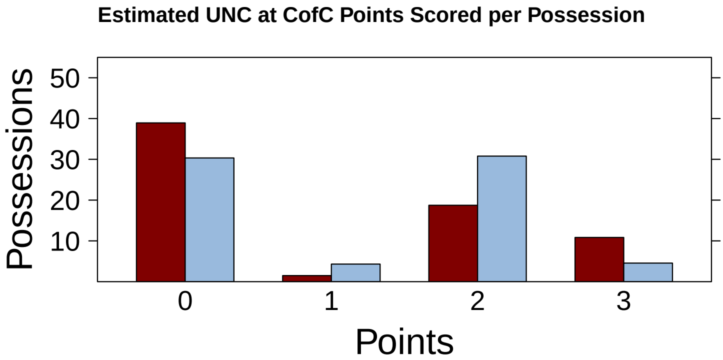
$$\log\left(\frac{\pi_i}{\pi_0}\right) = \alpha + \beta_0(\text{home}) + \beta_i + \beta_j, \text{ for } i = 1, 2, 3$$

- β_i: rating for team i
- home$=\begin{cases} 1 & \text{if team } i \text{ is at home} \\ -1 & \text{if team } i \text{ is away} \\ 0 & \text{otherwise} \end{cases}$

Estimating Points Scored on Possessions

- Play-by-play data is scarce
- Data can be estimated using the box score
- For example, to estimate the number of zeros:
  - $0.97 \times$ FGA-FGM $+ 0.27 \times$ FTA-FTM $- 0.96 \times$ OR $+ 1.02 \times$ TO
- Similar models for ones, twos, and $\geq$ threes

Estimates from the Multinomial Model for 2010



- Pr(CofC Win) = 0.26 (Linear Model: 0.30)

Multinomial Model: Probabilities of Winning

- Estimated with simulation
- Assumptions:
  - Possessions are independent
  - Each team will have n offensive possessions
- For the desired number of simulations:
  1. Simulate n possessions using model probabilities
  2. Determine winner of game (ignore ties)
- Use results to estimate probability of winning

Answering Questions with the Multinomial Model

- Who is the favorite to win the tournament?
  - Duke, 21% (Linear Model: Kansas, 25%)
- How often does Kentucky make the Final Four?
  - 18% (Linear Model: 17%)
- Will St. Mary's advance to the Sweet 16?
  - 26% (Linear Model: 25%)

Model Comparison: ESPN Scores

- Earn $2^{r-1} \times 10$ points for rounds $r = 1, 2, \ldots, 6$
- Maximum of 1920 points possible

Season	Linear	Multinomial	Difference
2003	790	590	200
2004	740	810	-70
2005	1310	1450	-140
2006	730	670	60
2007	1010	730	280
2008	1480	1570	-90
2009	750	780	-30
Mean	973	943	30

- Multinomial model won 4 out of 7 tournaments

Model Comparison: Differing Picks

- What happens when models disagree?
- From 2003 to 2010 (1^{st} round), models disagreed 36 times
- Linear model selected 22 correctly ($\hat{\pi} = 22/36 = 61\%$)
- Multinomial model selected 14 correctly (39%)
- 95% CI for π: (43%, 77%)

Future Work

- Calculate confidence intervals
- Logistic regression/Markov chain (LRMC) model comparison
- Estimate model prediction error

References

- Ken Pomeroy, *Stats Explained*, http://kenpom.com/blog/index.php/C24/P5/
- Kvam, P. and J.S. Sokol, *A logistic regression/Markov chain model for NCAA basketball*, Naval Research Logistics 53, pp. 788-803