Estimating the Probability of Winning a College Basketball Game

Ryan J. Parker
ryan@basketballgeek.com

College of Charleston

March 21, 2010
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?
Essential for non-standard point systems
### Tournament Path Matters: Extreme Example

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Win Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Team A</td>
<td>51%</td>
</tr>
<tr>
<td>2</td>
<td>Team B</td>
<td>49%</td>
</tr>
<tr>
<td>3</td>
<td>Team C</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>Team D</td>
<td>0%</td>
</tr>
</tbody>
</table>

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%
**Efficiency**: mean number of points scored per possession
- Removes effect of pace on a team’s points scored and allowed
- Possessions estimated with $\text{FGA} - \text{OR} + \text{TO} + 0.475 \times \text{FTA}$
$D_{ij}$: difference in team $i$’s and team $j$’s efficiency

Linear regression where we assume

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

$\beta_i$: 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

Histogram of residuals

Standardized Residuals

Normal Q–Q Plot

Sample Quantiles

Theoretical Quantiles
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
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:**

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

- $\beta_i$: rating for team $i$
- $\text{home} = \begin{cases} 1 & \text{if team } i \text{ is at home} \\ -1 & \text{if team } i \text{ is away} \\ 0 & \text{otherwise} \end{cases}$
Play-by-play data is scarce
Data can be estimated using the box score
For example, to estimate the number of zeros:
- \[ 0.97 \times \text{FGA-FGM} + 0.27 \times \text{FTA-FTM} - 0.96 \times \text{OR} + 1.02 \times \text{TO} \]
Similar models for ones, twos, and \( \geq \) threes
Estimates from the Multinomial Model for 2010

Estimated UNC at CofC Points Scored per Possession

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

<table>
<thead>
<tr>
<th>Season</th>
<th>Linear</th>
<th>Multinomial</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>790</td>
<td>590</td>
<td>200</td>
</tr>
<tr>
<td>2004</td>
<td>740</td>
<td>810</td>
<td>-70</td>
</tr>
<tr>
<td>2005</td>
<td>1310</td>
<td>1450</td>
<td>-140</td>
</tr>
<tr>
<td>2006</td>
<td>730</td>
<td>670</td>
<td>60</td>
</tr>
<tr>
<td>2007</td>
<td>1010</td>
<td>730</td>
<td>280</td>
</tr>
<tr>
<td>2008</td>
<td>1480</td>
<td>1570</td>
<td>-90</td>
</tr>
<tr>
<td>2009</td>
<td>750</td>
<td>780</td>
<td>-30</td>
</tr>
<tr>
<td>Mean</td>
<td>973</td>
<td>943</td>
<td>30</td>
</tr>
</tbody>
</table>

- Multinomial model won 4 out of 7 tournaments
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 $\pi$: (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/