Applying Multilevel Models to Sports Rating Problems

Ryan J. Parker

College of Charleston

August 14, 2009

Why Rating Matters (to me)

Who Cares About Ranking?

What is a Multilevel Model?

- **Motivation**: modeling data structured into groups
  - **Academics**: Students grouped by the school they attend
  - **Basketball**: Shots grouped by the player that attempted them
- Regression where coefficients are modeled
- These models give us a handle on variation between groups
- Can better estimate the effects of groups with small samples

Multilevel Model Formulation

- Consider a linear regression model that uses indicator variables to estimate group effects for $j = 1, \ldots, J$ groups using $i = 1, \ldots, n$ observations:

$$y_i \sim \mathrm{N}(\alpha_{j[i]}, \sigma_y^2)$$

- The multilevel model formulation:

$$\begin{aligned} y_i &\sim \mathrm{N}(\alpha_{j[i]}, \sigma_y^2) \\ \alpha_j &\sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2) \end{aligned}$$

- Group effects are assumed to come from a normal distribution, with the mean μ_α and variance σ_α^2 estimated from the data
- $j[i]$ is the index for the group associated with observation i

Classical Regression versus Multilevel Modeling

- **Goal**: Predict free throw shooting percentages of NBA centers
- Classical logistic regression model:

$$\begin{aligned} y_i &\sim \texttt{Binomial}(n_i, \theta_i) \\ \texttt{logit}(\theta_i) &= \texttt{log}\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_i \end{aligned}$$

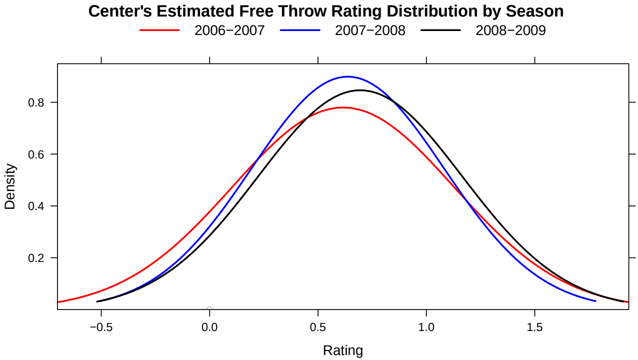
- Multilevel logistic regression model:

$$\begin{aligned} y_i &\sim \texttt{Binomial}(n_i, \theta_i) \\ \texttt{logit}(\theta_i) &= \beta_i \\ \beta_i &\sim \texttt{N}(\mu_\beta, \sigma_\beta^2) \end{aligned}$$

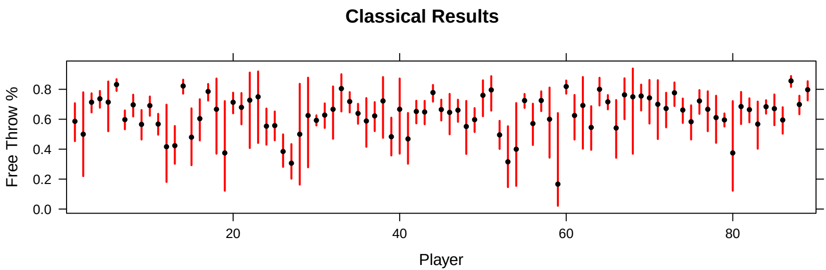
- Models the probability that player i makes a free throw attempt
- Data from three NBA regular seasons (2006-2007 to 2008-2009)

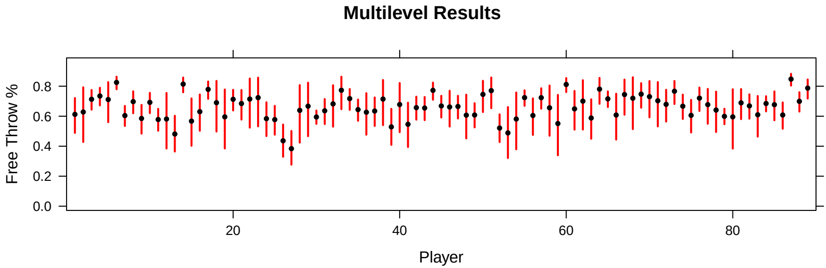
Classical vs Multilevel: Fit Details

- Classical fit estimates some $\beta_i = \infty$ or $-\infty$
- Multilevel fit estimates:
  2006-2007: $\hat{\mu}_\beta = 0.616$, $\hat{\sigma}_\beta = 0.512$, 95% CI: $(40.4\%, 83.5\%)$
  2007-2008: $\hat{\mu}_\beta = 0.637$, $\hat{\sigma}_\beta = 0.444$, 95% CI: $(44.2\%, 81.9\%)$
  2008-2009: $\hat{\mu}_\beta = 0.695$, $\hat{\sigma}_\beta = 0.471$, 95% CI: $(44.3\%, 83.5\%)$



Classical vs Multilevel: Fit Comparison (2008-2009)





Classical vs Multilevel: Predictions

Gazing into the crystal ball:

		Classical		Multilevel		Difference	
Model	**Actual**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
2007	2008	9.1%	13.9%	7.0%	10.2%	2.1%	3.7%
2008	2009	8.2%	12.6%	7.7%	11.7%	0.6%	0.9%
2007	2009	8.4%	12.7%	7.2%	10.1%	1.3%	0.9%
2007	08+09	7.5%	11.8%	5.5%	8.0%	2.0%	3.8%
07+08	2009	7.7%	11.9%	7.1%	10.8%	0.6%	1.1%

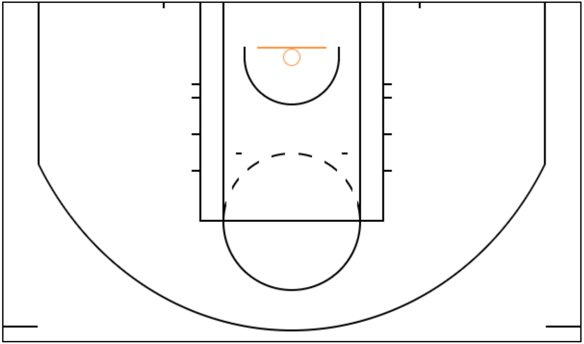
Minimum 10 attempts in ***Actual*** *data set*

MAE: Mean Absolute Error = $\frac{1}{n}\sum_{i=1}^{n}|predicted_i - actual_i|$

RMSE: Root Mean Squared Error = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(predicted_i - actual_i)^2}$

NBA Example: 3pt Shooting

- **Goal**: Predict a player's future 3pt field goal percentage
- What is home court advantage worth?
- What is the difference between regular and corner 3pt shots?
- Data from seven NBA regular seasons (2002-2003 to 2008-2009)



3pt Shooting: Ability by Position

- Basic multilevel model for each position:

$$\begin{aligned} y_i &\sim \texttt{Binomial}(n_i, \theta_i) \\ \texttt{logit}(\theta_i) &= \beta_i \\ \beta_i &\sim \texttt{N}(\mu_\beta, \sigma^2_\beta) \end{aligned}$$

- Estimates by position:

	μ_β	σ^2_β	**Mean 3FG%**	**95% CI**
PG	-0.586	0.127	35.8%	(30.3%, 41.7%)
SG	-0.547	0.121	36.7%	(31.3%, 42.3%)
SF	-0.595	0.120	35.5%	(30.3%, 41.1%)
PF+C	-0.754	0.186	32.0%	(24.6%, 40.4%)
All	-0.595	0.130	35.5%	(29.9%, 41.6%)

- PF+C only group statistically significant from the others

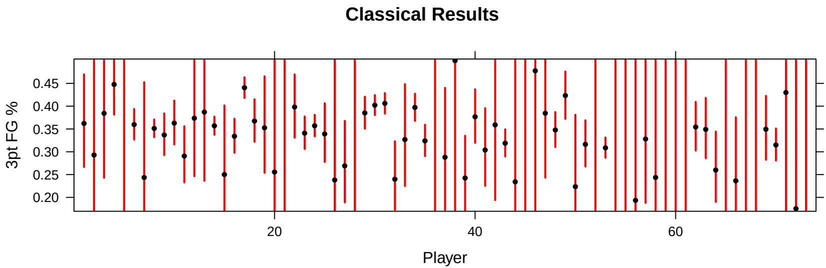
3pt Shooting: Model for HCA and Shot Location

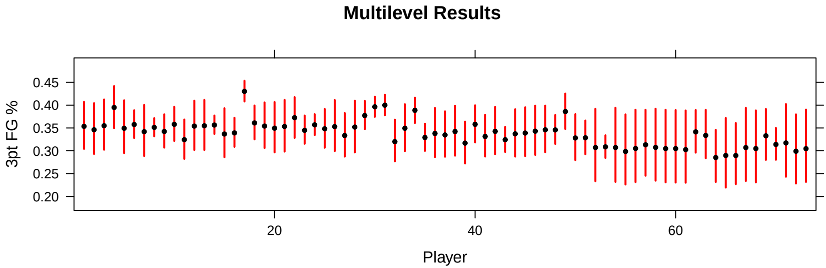
- Model for home court advantage and shot location:

$$
\begin{aligned}
y_i &\sim \texttt{Binomial}(n_i, \theta_i) \\
\texttt{logit}(\theta_i) &= \beta_i + \gamma(\text{home}) + \delta(\text{corner3}) \\
\beta_i &\sim \texttt{N}(\mu_\beta, \sigma_\beta^2)
\end{aligned}
$$

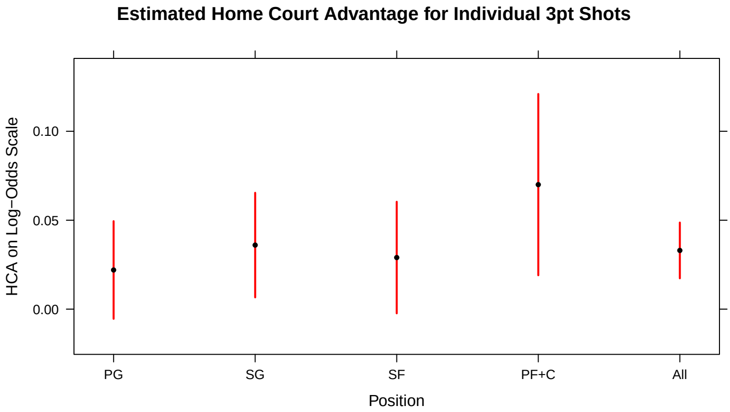
- **home**: 1 if shot attempt taken at home, 0 otherwise
  γ measures home court advantage
- **corner3**: 1 if shot attempt taken from the corner, 0 otherwise
  δ measures difference between regular and corner 3pt shots

3pt Shooting: Player Ratings



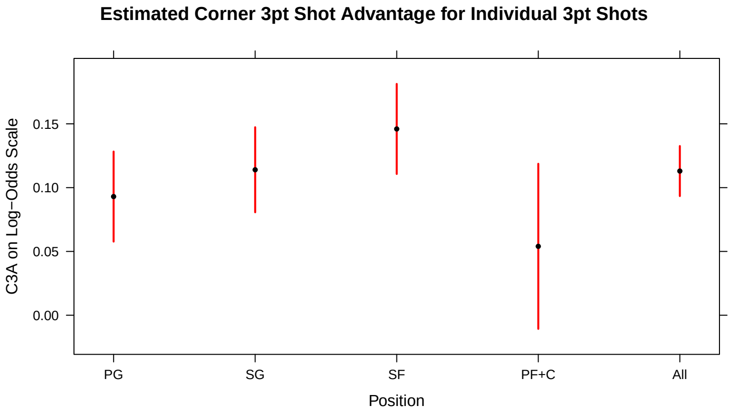


3pt Shooting: Home Court Advantage



Difference of $\sim$3 points per 100 shot attempts

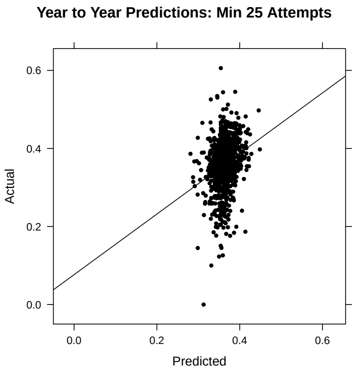
3pt Shooting: Regular versus Corner 3pt Shots



Difference of ∼9 points per 100 shot attempts

3pt Shooting: Year to Year Predictions

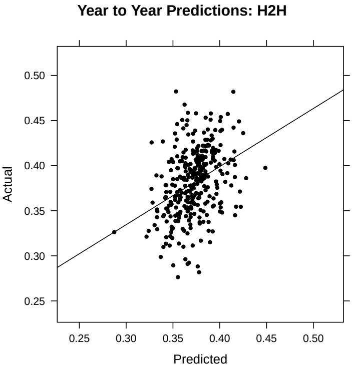
MAE	RMSE	Mean	SD	95% CI
4.6%	6.2%	0.5%	6.1%	(-11.6%, 12.5%)

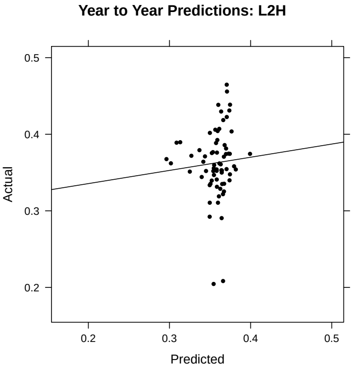


3pt Shooting: Predicting Different Player Groups

Group	MAE	RMSE	Mean	SD	95% CI
High to High	2.9%	3.7%	-0.9%	3.6%	(-8.0%, 6.2%)
Low to High	3.6%	4.7%	-0.5%	4.8%	(-9.8%, 8.8%)

High to High: > 150 attempts both years (n=315), **Low to High**: < 50 and > 100 attempts year to year (n=69)





The Classical Bradley-Terry Model

- Logistic regression model for paired comparisons
- Examples of this type of data:
  - **Product Prefs**: Bud Light or Miller Lite? Miller Lite or Coors Light?
  - **Tennis**: Agassi or Federer? Federer or Roddick?
- When i is compared to j:

$$\texttt{logit}(\theta_{ij}) = \texttt{log}\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) = \beta_i - \beta_j$$

- To estimate the probability that i is preferred to j:

$$\hat{\theta}_{ij} = \texttt{logit}^{-1}(\hat{\beta}_i - \hat{\beta}_j) = \texttt{exp}(\hat{\beta}_i - \hat{\beta}_j)/(1 + \texttt{exp}(\hat{\beta}_i - \hat{\beta}_j))$$

The Hierarchical Bradley-Terry Model

- Model the ratings β_t
- Hierarchical formulation:

$$\begin{aligned} Y_{ij} &\sim \texttt{Bernoulli}(\theta_{ij}) \\ \texttt{logit}(\theta_{ij}) &= \beta_i - \beta_j \\ \beta_t &\sim \texttt{N}(0, \sigma^2_\beta) \\ \sigma^2_\beta &\sim \texttt{Gamma}(a, b),\ a \text{ and } b \text{ known} \end{aligned}$$

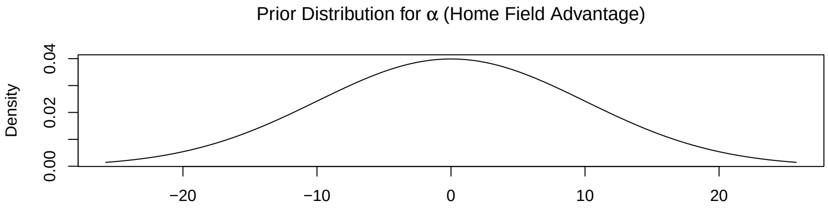
2008 CFB Example: Hierarchical Formulation

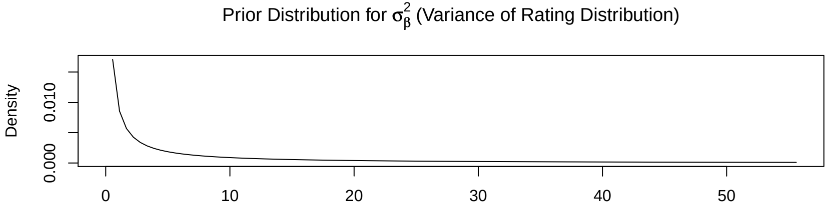
- **Motivation**: Classical Bradley-Terry model estimates some ratings to be ∞ or $-\infty$
- Hierarchical Bradley-Terry model for college football:

$$
\begin{aligned}
Y_{ij} &\sim \texttt{Bernoulli}(\theta_{ij}) \\
\texttt{logit}(\theta_{ij}) &= \alpha(\text{home}) + \beta_i - \beta_j \\
\alpha &\sim \texttt{N}(\mu_\alpha, \sigma_\alpha^2),\ \mu_\alpha \text{ and } \sigma_\alpha^2 \text{ known} \\
\beta_t &\sim \texttt{N}(0, \sigma_\beta^2) \\
\sigma_\beta^2 &\sim \texttt{Gamma}(a, b),\ a \text{ and } b \text{ known}
\end{aligned}
$$

- **Prior Params**: $\mu_\alpha = 0$ and $\sigma_\alpha^2 = 100$; $a = 0.01$ and $b = 100$.
- home = 1 if i is at home; -1 if i is away; 0 if neutral site

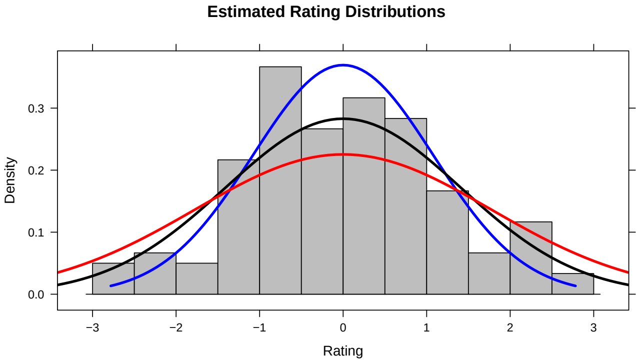
2008 CFB Example: Prior Distributions



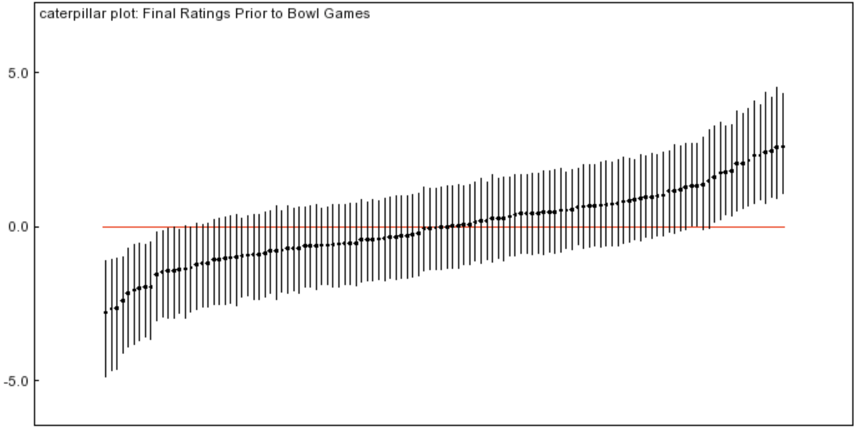


2008 CFB Example: Hierarchical Fit

- **Estimates**:
  - $\hat{\alpha} = 0.50$ and 95% credible interval for α: $(0.30, 0.71)$
    Home Win%: 62% or $(57\%, 67\%)$ for evenly matched teams
  - $\hat{\sigma}_\beta = 1.41$ and 95% credible interval for σ_β: $(1.08, 1.77)$



2008 CFB Example: Hierarchical Fit (cont)



2008 CFB Example: Hierarchical Fit (cont)

Top Rated Team: Oklahoma

Team	Pr(Oklahoma Better)	Pr(Oklahoma Wins)
Utah	51.3%	50.9%
Texas	55.0%	53.6%
Boise State	56.3%	54.9%
Florida	59.9%	57.1%
Texas Tech	59.7%	57.2%
Alabama	65.0%	60.9%
Southern California	68.1%	63.3%
Penn State	69.1%	63.7%
Ohio State	75.9%	69.1%
Cincinnati	81.6%	69.7%

Conclusion

- Multilevel models help rate groups with varying sample sizes
- Predictions are better than a naive classical approach
- Care must be taken when examining specific predictions
- Hierarchical models help us fit a variety of formulations

References

1. Data and code used in the making of this presentation available at `http://www.basketballgeek.com/downloads/rw09.zip`
2. I.Ntzoufras: `Bayesian Modeling Using WinBUGS`
3. A.Gelman, J.Hill: `Data Analysis Using Regression and Multilevel/Hierarchical Models`
4. I.Witten, E.Frank: `Data Mining`
5. A.Agresti: `An Intro to Categorical Data Analysis`
6. CFB data from *James Howell's College Football Scores*: `http://homepages.cae.wisc.edu/~dwilson/rsfc/history/howell/`