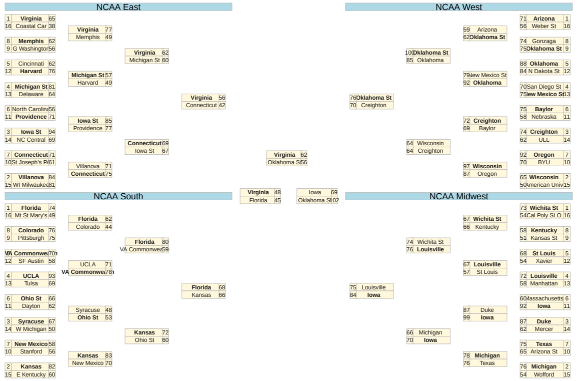
Quantile Bracketology

Roger Koenker

University of Illinois, Urbana-Champaign

Statistics Brown Bag: 20 February 2015



Four Prior Stipulations

Surgeon General's Warning

Gambling can be dangerous for your wealth.

Casey Stengel's Warning

Never make predictions, especially about the future.

Colin Mallow's Warning

I try not to think about this too much; it is too much fun.

My Warning

I know nothing about basketball,
this is pure exploratory data analysis.

Motivation

For the second year Kaggle (kaggle.com) is running a competition sponsored by HP to predict the outcome of the NCAA Men's Basketball Tournament. Data is provided for the last 30 years of college basketball. Entrants predict the probabilities of every possible match-up and entries are scored by the (logistic) loss function,

$$L(y,p) = -n^{-1}\sum_{i=1}^{n}(y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i))$$

where y_i is the binary outcome of game i of the tournament, and $\hat{p}_i$ is the entrants predicted probability for game i. The winning entrant gets \$10K, second place gets \$5K.

The Classical Binary Paired Comparison Model

Let Y_{ijg} denote the score of team i playing team j in game g and suppose:

$$\Lambda(P\{Y_{ijg} = 1\}) = \alpha_i - \alpha_j + \gamma D_g$$

where Λ is a specified link function, say logistic, the α parameters are ratings for teams i and j, and $D_g = I(\text{game } g \text{ is played on team } i\text{'s home court})$, so γ denotes the home court advantage. This model is identifiable (estimable) provided that there is sufficient overlap in scheduling of the observed games. There needs to be some reasonable amount of inter-conference competition. A good reference is H.A. David (1988) *The Method of Paired Comparisons*.

Critique of the Binary Paired Comparison Model

- Binary response sacrifices information on the winning margin.
- Ignores distinction between offensive and defensive capability.

The Mean Paired Comparison Model

Let Y_{ijg} denote the score of team i playing team j in game g and suppose:

$$EY_{ijg} = \alpha_i - \delta_j + \gamma D_g$$

Now we can estimate offensive and defensive ratings for each team, by least squares.

Critique of the Mean Paired Comparison Model

- Presumes Gaussian “errors,” so extreme scores (blowouts) can exert “too much” influence on ratings,
- Presumes homoscedastic “error” so all (games) scores have the same variability.

A Quantilesque Paired Comparison Model

Suppose instead of postulating a model for mean scores we posit a model for the quantiles of scores:

$$Q_{Y_{ijg}}(\tau) = \alpha_i(\tau) - \delta_j(\tau) + \gamma(\tau) D_g$$

- Median version ($\tau = 1/2$) is quite similar to mean model,
- Except that it is less sensitive to extreme scores,
- For general τ we permit much richer class of rankings
- Some teams can be very consistent others very erratic
- Teams can have different shapes for their ratings functions
- Mean model is nested: $Q_{Y_{ijg}}(\tau) = \alpha_i - \delta_j + \gamma D_g + \Lambda^{-1}(\tau)$

Prediction in the QPCM

Suppose teams i and j meet at a neutral site, the result is modeled by the quantile functions for the two scores:

$$(Q_{Y_{ig}}(\tau), Q_{Y_{jg}}(\tau)) = (\alpha_i(\tau) - \delta_j(\tau), \alpha_j(\tau) - \delta_i(\tau))$$

We can simulate the probability of team i winning by Δ.

$$\pi_{ij} = P(Q_{Y_{ig}}(U) > Q_{Y_{jg}}(V) + \Delta).$$

where U and V are (independent??) uniforms, provided we know the α’s and δ’s. We’ll return to the dubious independence assumption.

Estimation of the QPCM

Estimation is just a (very sparse) quantile regression problem:

$$\min_{(\alpha,\delta,\gamma)} \sum_g \rho_\tau(y_{ig} - \alpha_i + \delta_j - \gamma D_{ig}) + \rho_\tau(y_{jg} - \alpha_j + \delta_i - \gamma D_{jg})$$

or,

$$\min_\theta \|y - X\theta\|_\tau,$$

where $\|u\|_\tau \equiv \sum \rho_\tau(u_i) \equiv \sum u_i(\tau - I(u_i < 0))$, $y = (y_i, y_j)$ denotes a stacked vector of scores, $\theta = (\alpha, \delta, \gamma)$ and

$$X = \begin{bmatrix} H & -A & D_i \\ A & -H & D_j \end{bmatrix}$$

with $H_{g,i} = 1$ if team i is the Home team of game g, and $= 0$ otherwise, $A_{g,j} = 1$ if j is the Away team of game j, and $= 0$ otherwise, and D_i and D_j denote the home court indicators. No row of X has more than 3 non-zero entries!

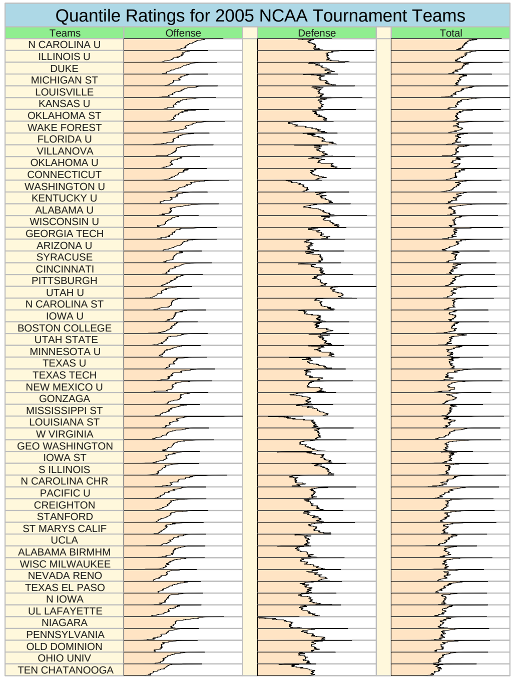
Quantile Regression Bracketology: Estimation

For last year's Kaggle competition I estimated a model based on 5362 games involving 350 teams, on grid of 200 equality spaced τ's. The design matrix X was therefore 10724 by 702 and is 99.5% zeros. It takes about a minute to do this on my MacPro desktop machine.

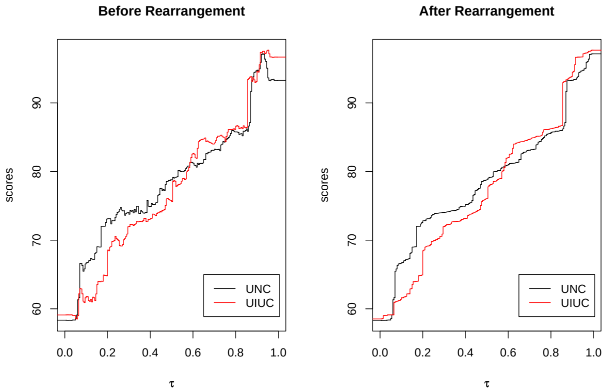
For our earlier JBES (2010) paper, Gib Bassett and I estimated the model on a sample of 2940 games involving 232 Division I NCAA college basketball teams for the 2004-05 regular season. The estimated model was then used to predict the outcomes of the 2005 NCAA basketball tournament.

This happened to be a season in which UIUC did well, losing to the University of North Carolina only in the final game of the tournament.

Estimation Results for 2005

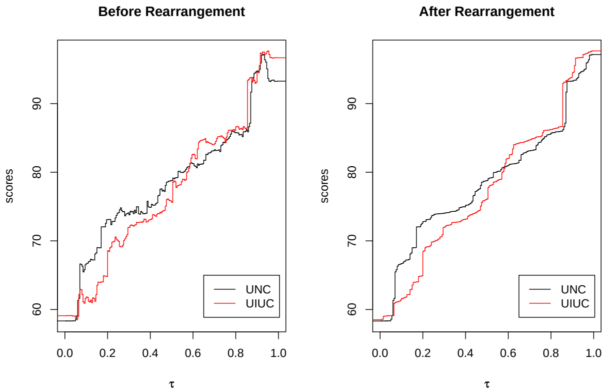


Predicting the 2005 Final Game UIUC v. UNC



Estimated quantile functions for scores with and without monotonization à la Chernozhukov, Fernandez-Val, and Galichon (2006).

Predicting the Final Game UIUC v. UNC



UNC is predicted to win when the game is low scoring, UIUC has the advantage when the game is high scoring.

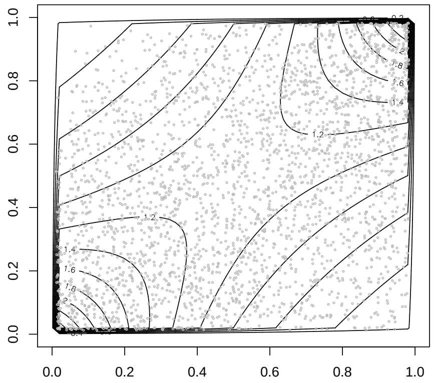
Are Within Game Scores Really Independent?

One way to explore possible dependence of scores is to consider,

$$\hat{u}_{kg} = \int_0^1 I(y_{kg} \leqslant \hat{Q}_{kg}(\tau))d\tau, \quad k = i, j.$$

These quantities are something like QR residuals, they purport to tell us what quantile of the conditional distribution a particular realized score fell onto. Marginally, by construction they are approximately uniform. So plotting these pairs suggests estimating a copula function.

Within Game Score Dependence



The estimated Frank copula parameter, $\hat{\theta} = 2.52$, is highly significant, confirming the highly significant Kendall rank correlation of .27, and indicating a positive association between pairs of scores.

The Random Coefficient Score Model

Another interpretation of the QPCM is that scores are generated as:

$$Y_{ig} = \alpha_i(U) - \delta_j(U) + \gamma(U)D_{ig},$$

$$Y_{jg} = \alpha_j(V) - \delta_i(V) + \gamma(V)D_{jg},$$

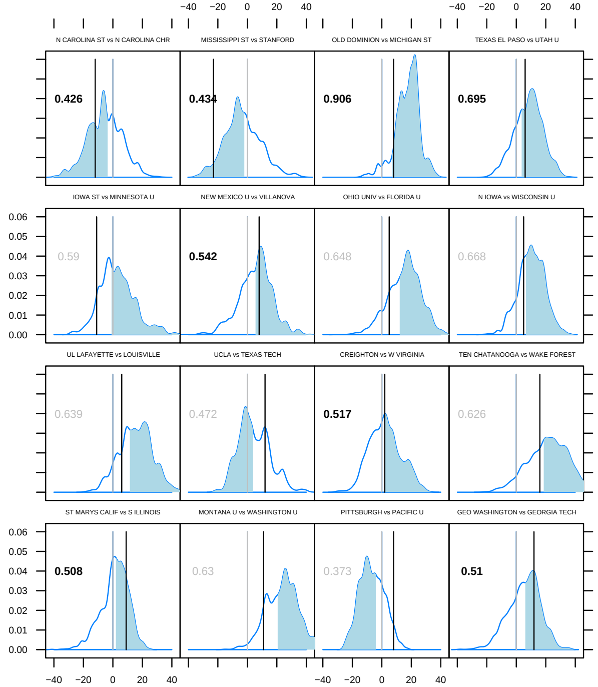
where U and V are uniform random variables on $[0, 1]$. Drawing U and V from our estimated copula yields a mechanism for simulating the predictive density for games between teams i and j.

Simulating the Point Spreads

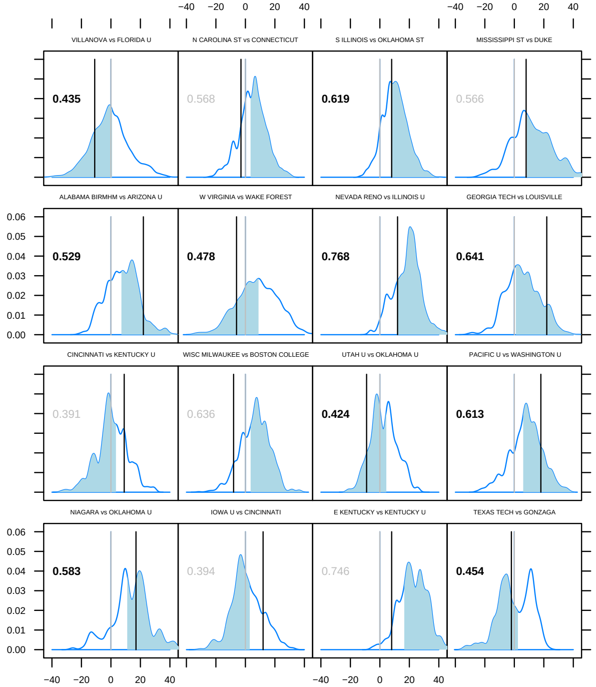
For the games of the 2005 NCAA Tournament we simulated 10,000 realizations of the point spread $Y_{ig} - Y_{jg}$ for each game:

- Using *all* the games prior to the round of the game for estimation,
- Treating the tournament venues as neutral sites,
- Estimating densities using standard kernel method in R,
- Vertical grey lines to indicate a tie score,
- Shaded blue region to indicate Las Vegas pointspread,
- Vertical black line to indicate the realized pointspread.

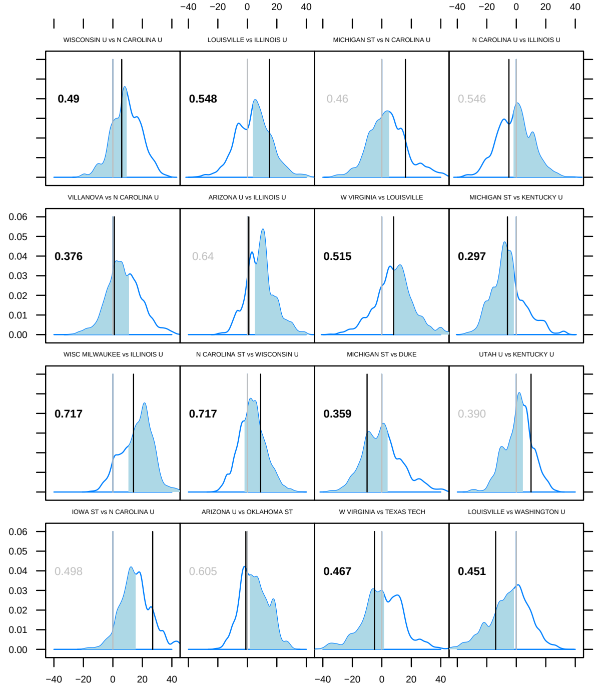
Predictive Densities for 2005 Tournament



Predictive Densities for 2005 Tournament



Predictive Densities for 2005 Tournament



Quantile Regression Bracketology: Survival Curves

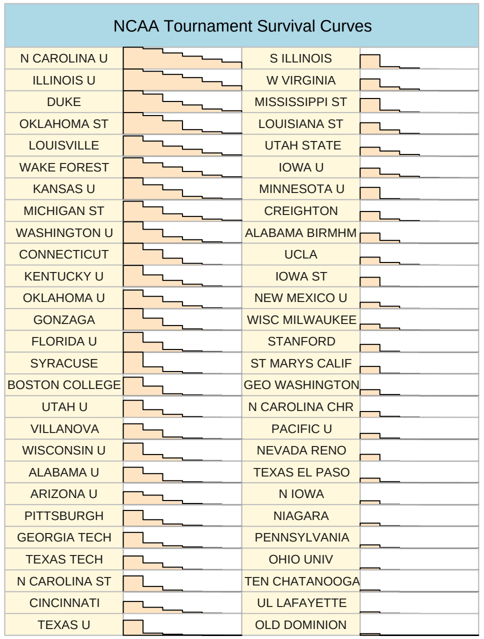
Predicting pointspreads is useful for betting, but it doesn't help you fill in the tournament bracket, which is a much more popular form of College Basketball betting. For this, you need to estimate the likelihood of various tournament pairings:

- Given the estimated model,
- We estimated 1000 realizations of the tournament,
- Starting from the original tournament pairings.

Predicted Tournament Performance

Criterion	UNC	UIUC	Duke
$\mathbb{E}$ Exit Round	4.025	3.905	2.953
$\mathbb{P}$ Champion	0.318	0.233	0.083

Quantile Regression Bracketology: Survival Curves



Betting on the Pointspread

How well would we have done betting on the Las Vegas pointspreads in the 48 tournament games we have illustrated?

- Bet on the team with best probability of beating the pointspread,
- In 27 out of 47 games we have bet correctly,
- One game was a push so the money bet is refunded.
- It costs \$110 to place a \$100 bet, so
- We have an expected gain of \$10.54 on each \$100 bet, with $p = 27/47$, $\mathbb{E}G = 100p - 110(1 - p) = 10.54$.

Betting on the Over/Under

It is also possible to bet on the sum of the scores rather than their difference.

- We compute predictive densities for the score totals,
- Again, there are posted Las Vegas "point totals,"
- We employ the same betting strategy,
- Coincidently, we also get 27 out of 47 correct.

Should We Quit Our Day Jobs?

Probably not:

- 48 games is a rather small sample, but
- Better than picking up nickels in front of a steamroller,
- There are many possible refinements:
  - Shrinkage to control variability of the profligate model specification,
  - Weighting to accentuate the import of most recent games,
  - Introduction of prior season performance
  - Introduction of other covariates
- But evidence for the Hayek hypothesis that aggregation of market bets yields accurate probability assessment, is rather weak.

Kaggle Round Two?

What should I do differently?

- Pay (more) attention to the copula model!
- Use several years prior data?
- Penalties/Shrinkage?
- Other ideas?

Slides and an R package for all of this will be available from my webpages.