



Taylor & Francis
Taylor & Francis Group

March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis

Authors(s): Roger Koenker and Gilbert W. Bassett Jr.

Source: *Journal of Business & Economic Statistics*, Vol. 28, No. 1 (January 2010), pp. 26-35

Published by: Taylor & Francis, Ltd. on behalf of American Statistical Association

Stable URL: <http://www.jstor.org/stable/27799112>

Accessed: 24-03-2016 01:50 UTC

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

<http://about.jstor.org/terms>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



American Statistical Association, Taylor & Francis, Ltd. are collaborating with JSTOR to digitize, preserve and extend access to *Journal of Business & Economic Statistics*

<http://www.jstor.org>

March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis

Roger KOENKER

Department of Economics, University of Illinois at Urbana–Champaign, Urbana, IL 61801 (rkoenker@uiuc.edu)

Gilbert W. BASSETT JR.

Department of Finance, University of Illinois at Chicago, Chicago, IL 60607

A quantile regression variant of the classical paired comparison model of mean ratings is proposed. The model is estimated using data for the regular 2004–2005 U.S. college basketball season, and evaluated based on predictive performance for the 2005 National Collegiate Athletic Association (NCAA) basketball tournament. Rather than basing predictions entirely on conditional mean estimates produced by classical least-squares paired comparison methods, the proposed methods produce predictive densities that can be used to evaluate point spread and over/under gambling opportunities. Mildly favorable betting opportunities are revealed. More generally, the proposed methods offer a flexible approach to conditional density forecasting for a broad class of applications. An electronic supplement containing predictive densities for point totals and team offensive and defensive ratings is available from the *JBES* website.

KEY WORDS: Paired comparison; Quantile regression.

“Though this be madness, yet there is method in’t.” (Hamlet [II, ii])

1. INTRODUCTION

To concentrate the mind, imagine yourself on the evening of Selection Sunday with a decent laptop and a good internet connection struggling with the task of “filling out the NCAA tournament bracket,” that is, picking the winners of the games of the impending single elimination tournament. Although prior season performance of the teams is readily available from internet sources, a more challenging task is to find a plausible model capable of making credible predictions. Ideally, we would like a model that will predict for any particular pairing of teams the joint density of their final scores. From such a predictive density one can then design betting strategies based on point spreads, odds, the over/under, or other gambling opportunities. Conventional paired comparison models deliver, at best, an estimate of a pair of mean scores, under the maintained hypothesis that there is a homogeneous bivariate normal density centered at this estimate. Often, however, only the mean of the score difference is estimated, or even more simply, a predicted probability of a victory based on binary data on won–lost records.

Adapting quantile regression methods to the paired comparison framework, we describe a model that is capable of delivering estimates of this conditional joint density. The approach is illustrated by estimating the model for the 2004–2005 National Collegiate Athletic Association (NCAA) college basketball season. Evaluation of the performance of the methods is based on out-of-sample predictive performance for the 2005 NCAA Division I tournament.

There is an extensive literature on sports betting, or what is known more euphemistically in economics as wagering markets. For a valuable survey see Sauer (1998). It is often claimed that such markets reveal something important about how het-

erogeneous probabilistic information about athletic contests gets baked into the Hayek (1945) cakes of efficient odds, point spreads, and yet more esoteric gambles. A multitude of published studies documented small discrepancies that seem to undercut the “market efficiency of sports betting.” But the economist’s normal skepticism entitles us to ask: Can this just be the flip side of the familiar publication bias worry? Betting strategies that win, but fail to overcome the vigorish, get published; strategies that succeed go directly to Vegas.

Our predictions clearly reveal favorable betting opportunities well in excess of the prevailing vigorish. Instead of rushing off to Las Vegas, we prefer to disseminate the methods, which have a broad range of potential applications in other statistical domains where paired comparison data arise. Rarely are the convenient simplifying assumptions of the classical paired comparison methods easily justified. When they are violated, there are potential gains from estimating more flexible models like those proposed in the following.

Paired comparison data arise not only in sports, but in many other settings. Multiple treatments are evaluated with paired comparisons in clinical trials, consumer product testing, evaluation of expert testimony, page ranks for web pages, educational testing, and a variety of other contexts. More generally, the methods suggested here illustrate a flexible approach to the estimation of conditional densities that have applications in many other settings beyond the paired comparison model where effective forecasting requires more than a conditional mean prediction.

After a brief critique of classical paired comparison methods in Section 2, we introduce our model in Section 3, describe estimation methods in Section 4, and evaluate the performance of the methods in Sections 5 and 6.

© 2010 American Statistical Association
Journal of Business & Economic Statistics
January 2010, Vol. 28, No. 1
DOI: 10.1198/jbes.2009.07093

2. WHAT WE ARE NOT PROPOSING TO DO

Two hundred years of statistical inertia might suggest that we begin by considering a paired comparison model for expected scores that looks like this:

$$EY_{ig} = \alpha_i - \delta_j + \gamma D_{ig}, \quad (2.1)$$

where Y_{ig} denotes the score of team i in game g against opponent j . The parameter α_i may be interpreted as an offensive rating of team i , δ_j is a defensive rating of team j , and γ denotes a generic home court advantage, if any, so D_{ig} takes the value 1 if game g is played on team i 's home court, and takes the value 0 otherwise. Least-squares estimation of this conditional mean model will presumably use all of the results prior to the tournament. Each game will contribute two observations and we estimate a vector of offensive and defensive ratings for each of the m teams that were potential candidates for the tournament. We employ a version of this model as a point of comparison for our forecasting evaluation. See David (1988) for a definitive treatment of the classical theory of paired comparisons.

Estimation of this model by conventional least-squares methods raises several immediate concerns:

- (i) The model assumes that offensive and defensive performances of teams differ only in location while variation around these mean values is symmetric, constant across teams, and approximately Gaussian in shape. Thus, team effects are confined to the shifting location of the score densities, but since they have no effect on the scale or shape of these densities the model cannot capture the possibilities that some teams are more consistent than others or exhibit some form of asymmetry in their performance.
- (ii) Violations of Gaussian assumptions can introduce serious ratings anomalies, see, e.g., Bassett (1997), since a few games with extreme scoring can have undue influence on estimated parameters.
- (iii) The estimation of such a large number of parameters with relatively few observations is questionable; typically we might expect to have about 200 teams, thus 400 parameters, and about 3000 observations. Identification requires that there are not isolated groups of teams that never play common opponents, but even when this minimal condition is satisfied it may be advantageous to consider regularization schemes that introduce some form of "prior information."
- (iv) One may wish to question the independence assumption underlying an ordinary least-squares estimation of Equation (2.1). Teams may be thought to have "momentum" over the course of the season, introducing positive dependence in their performance. Perhaps even more plausibly, there may be dependence between the pair of scores for each game.
- (v) Finally, we may wish to predict the outcomes of games to evaluate performance of the model, with respect to potential gambling opportunities, purely as a matter of academic curiosity, of course, but these opportunities may require more than mean forecasts.

It may seem to be expecting a lot to resolve all of these issues in one brief article, but then *expecting* isn't our game.

3. A QUANTILE REGRESSION PAIRED COMPARISON MODEL

Rather than modeling expected scores according to Equation (2.1), suppose instead that we model conditional quantiles in an analogous fashion,

$$Q_{Y_{ig}}(\tau) = \alpha_i(\tau) - \delta_j(\tau) + \gamma(\tau)D_{ig}. \quad (3.1)$$

We maintain the assumption that there is a simple additive effect model in which now quantiles of team i 's score against team j are determined by the difference in their offensive and defensive ratings shifted by a home court advantage, $\gamma(\tau)$. However, these ratings may now be τ -dependent. If we fix $\tau = \frac{1}{2}$ for a moment and compare Equations (2.1) and (3.1) we are simply replacing a model for the conditional mean with one for the conditional median. The latter has a significant advantage, however, since it will be less sensitive to the tail behavior of the underlying random variables representing scores, and consequently, will be less sensitive to observed outliers in scores when it comes time to estimate the model. We maintain throughout the assumption that games are independent realizations, but will explore the within-game dependence of scores using copula methods.

Like the mean model, the median model is a pure location shift model; no provision is yet made for differences in the variability of a team's performance, their consistency, if you will. One might also imagine other more subtle differences in the shape of teams' scoring distributions. The specification of Equation (3.1) allows both offensive and defensive performances of teams to vary in dispersion, symmetry, or more exotic shape characteristics. Of course, this increased flexibility comes at a price. We replace an already rather profligate model with several hundred parameters with an even more profligate one in which each former parameter is now a *function* mapping the unit interval into the real line. Before turning to our discussion of estimation of this model, we briefly describe how it might be used for prediction.

4. QUANTILE REGRESSION BRACKETOLOGY

One may well ask: Given such a complicated model how are we to make predictions from it? This question has a surprisingly simple answer, if we accept, for the moment, the working premise that given the ratings, scores are independent. Consider the problem of predicting the winner of a game between teams i and j at a neutral site. The model provides quantile functions for the two scores:

$$Q_{Y_{ig}}(\tau) = \alpha_i(\tau) - \delta_j(\tau)$$

and

$$Q_{Y_{jg}}(\tau) = \alpha_j(\tau) - \delta_i(\tau).$$

The score of such a game can thus be represented, under our independence assumption, by the pair of random variables, $(Q_{Y_{ig}}(U), Q_{Y_{jg}}(V))$ where U and V are independent standard uniform random variables. The probability of team i winning by some fixed margin Δ at a neutral site is thus,

$$\pi_{ij} = P(Q_{Y_{ig}}(U) > Q_{Y_{jg}}(V) + \Delta). \quad (4.1)$$

Given explicit forms for the α 's and δ 's, this probability is easily approximated by simulation methods. We defer the question of possible dependence of U and V until Section 6, where we will be able to bring some empirical evidence to bear upon it.

To predict winners we simply set $\Delta = 0$ as above, and choose team i if $\pi_{ij} > \frac{1}{2}$ and choose team j otherwise. Predicting exact scores is an obviously much more challenging task. But, in principle, the model, by specifying the joint distribution of the game's final scores, provides everything that is necessary.

5. ESTIMATION

Estimation of Equation (3.1) is most straightforward if we begin by considering unconstrained estimation of the model for a single quantile. For each game g we have a pair of scores (y_{ig}, y_{jg}) . Maintaining our working independence assumption we wish to solve,

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

where $\rho_\tau(u) = u \cdot (\tau - I(u < 0))$. The resulting estimator $\hat{\theta}(\tau) = (\hat{\alpha}(\tau), \hat{\delta}(\tau), \hat{\gamma}(\tau))$ consistently estimates the parameters of the conditional quantile in Equation (3.1) under the conditions discussed in Koenker (2005). In the present context, these conditions posit a sequence of estimation problems with a fixed configuration of teams and the number of games tending to infinity in such a way that the schedule maintains the full rank condition specified in A2(i) of theorem 4.1 of Koenker (2005), a requirement that necessitates some regular interleague play between the teams. Balancing the plausibility of such assumptions over longer time horizons with stationarity assumptions on model parameters requires careful consideration. Consistency does *not* require independence of the responses (y_{ig}, y_{jg}) , indeed, there is a large amount of literature on estimation of quantile regression models under dependent conditions, see, e.g., Koenker (2005, section 4.6), and the references cited therein. What is crucial is the validity of the hypothesized Equation (3.1), which ensures that the expected value of the objective function is minimized at $\theta = (\alpha, \delta, \gamma)$ and that we have some control over the severity of the dependence. The latter consideration is trivially assured by the m -decomposability condition of Portnoy (1991) since the observations are one-dependent under our assumption that observations are independent across games.

Suppose we have n games among m teams in the pretournament sample. Defining n by m matrix H with g th row having i th element one, and remaining elements zero, and n by m matrix A with g th row having the j th element equal to one, we can rewrite the problem in matrix notation as

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

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

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

Here the n vectors D_i and D_j are indicators for whether the g th game is a home game for teams i or j , respectively. Of course, some games are played at neutral sites early in the season and for these games the entries in both D vectors will be zero.

The dimensionality of the matrix X is somewhat alarming, but modern developments in sparse linear algebra make solving problems like the one specified above very easy. The algorithm used to compute $\hat{\theta}$ is the sparse version of the Frisch–Newton interior point method described in Portnoy and Koenker (1997) and Koenker and Ng (2005). The sparsity of the design matrix X is quite extreme for these paired comparison models: there are, at most, three nonzero elements in any row of the X matrix. A computer representation of the problem requires only the storage of these nonzero elements and their indexes. Also, as noted in Koenker and Ng (2005), the computational effort is roughly proportional to the number of nonzero elements, so estimation even over a grid of several hundred τ 's is quite quick, requiring only a few minutes. Estimation of the quantile regression model was carried out with the `quantreg` package of Koenker (2006) designed for the R environment. Estimation of the corresponding conditional mean model was done with the `SparseM` package of Koenker and Ng (2006) for the same environment. It is worth emphasizing that even the least-squares version of the model will be a very challenging estimation problem in the absence of sparse linear algebra given that we are estimating 464 parameters. Further details on the computations, including data and all software used to produce tables and figures, are available from <http://www.econ.uiuc.edu/~roger/research/bracketology/MM.html>.

6. TASTING THE PUDDING: THE 2004–2005 SEASON

We estimate Equation (3.1) using data on 2940 games involving 232 Division I NCAA teams from the 2004–2005 basketball season. These games all occurred on or before Selection Sunday, March 13, 2005. Predictions reported in the following are based on estimation of the model using games through the tournament round preceding the prediction. Thus, the final game between the University of North Carolina (UNC) and the University of Illinois at Urbana–Champaign (UIUC) prediction uses all of the tournament game data, except, of course, for the final game itself. The model is estimated on an equally spaced grid of $J = 199$ quantiles $\tau \in (0, 1)$. Thus, for each of the 232 teams, we have an estimated offensive and defensive rating function evaluated at J points. The electronic Appendix to the article provides a graphical representation of these estimates and some associated one-dimensional rankings. In addition, we estimate a “home court advantage” that varies from somewhat more than three points per game to somewhat less than two points, as τ varies from 0 to 1. This home court effect is set to zero for our predictive exercises since tournament games are played on neutral courts, just as for the pretournament games on neutral courts in the estimation stage.

Our estimation method treats games as independent realizations, and assumes moreover, that the two realized scores for each game are also independent. The latter assumption seems particularly implausible. To explore the possible dependence of within-game scores we rely on a quantile regression specific notion of “residuals.” Given a realized score y_{ig} for team i in game g we ask: At what estimated quantile does this score fall? More explicitly, we compute the pairs,

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

By construction, these two random variables will each be approximately uniformly distributed. We can regard the pair (u_{ig}, u_{jg}) as quantile regression “residuals” for the game g . See Koenker (2005, section 3.5.4) for further details and the link to the regression rank score statistics of Gutenbrunner and Jurčková (1992). Equation (3.1) can be interpreted by Koenker (2005, section 2.6) as a random coefficient model in which scores are generated by

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

and

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

where U and V are uniform random variables on $[0, 1]$. Under this model, the pair (u_{ig}, u_{jg}) are the natural estimates of the corresponding pair (U, V) .

Recall that for any bivariate distribution function $F_{X,Y}(x, y)$ with marginals $F_X(x)$ and $F_Y(y)$, we define the copula function

$$C(u, v) = F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v)).$$

The copula function may be interpreted as the joint distribution function of the random variables $U = F_X(X)$ and $V = F_Y(Y)$ with uniform marginals, and concisely represents the dependence between the original variables X and Y . Thus, potential dependence between scores within games can be explored by fitting copula models to the pairs (u_{ig}, u_{jg}) . Their scatterplot, appearing in Figure 1 for our sample of 2940 games, exhibits some clustering along the diagonal, indicating a weak positive dependence in the two scores; a slowdown of the pace of the game by one team lowers scores for both teams. Figure 1 superimposes contours of the one-parameter Frank copula estimated by maximum likelihood. The estimated copula parameter of 2.52 with a standard error of 0.12 is highly significant, reinforcing the implausibility of the independence assumption. The (Kendall) correlation of the pair of scores is 0.27, which is also highly significant and closely matches the value obtained by simulation from the estimated copula. It may be eventually possible to improve the efficiency of the estimation of the model by exploiting knowledge of this dependence—in the spirit of Zellner’s seemingly unrelated regression (SUR) model—but we

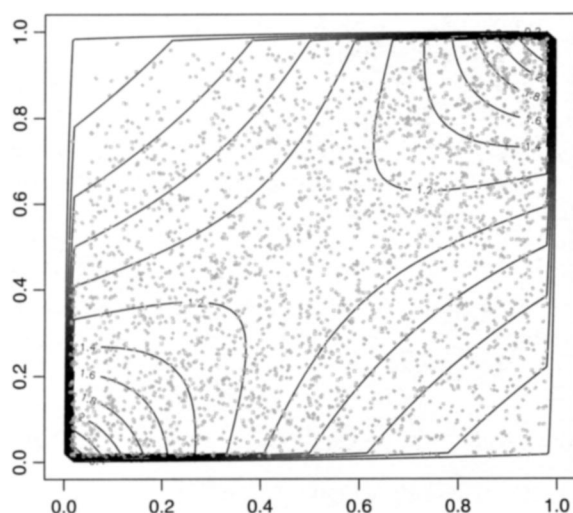


Figure 1. Estimated Frank copula density contours for final scores.

will not pursue this here. Consistency of the estimated ranking functions follows, as we note previously, from existing results, and this justifies the two-stage approach that we adopt. This dependence will be accounted for in our predictions where it plays an important role.

An extreme way to account for dependence in the within-game scores is to assume that both scores are generated by the same quantile “draw.” Recall that if a random variable X has distribution function F and associated quantile function $Q(u) = F^{-1}(u)$, then we can simulate realizations from X by generating uniform random variables, U , and computing $X = Q(U)$. This follows immediately from the fact that,

$$P(X \leq x) = P(Q(U) \leq x) = P(U \leq F(x)) = F(x).$$

Thus, for example, if we consider the final game of the 2005 NCAA tournament between UNC and UIUC, we can make predictions based on scoring outcomes of the form

$$(\hat{s}_i, \hat{s}_j) = (\hat{Q}_{ig}(\tau), \hat{Q}_{jg}(\tau)),$$

by simply replacing τ in the above expression by a draw from $U \sim U[0, 1]$. In the copula model this corresponds to all the mass of the copula concentrated on the 45% line of the unit square. In this case, we also have a linear conditional quantile model for score differences. For the classical paired comparison model the linearity of the conditional mean specification implies a corresponding model for score differences, but this is generally not the case for the quantile regression specification in Equation (3.1) unless we impose the condition of comonotonicity of the score pairs. We briefly consider this special case before turning to a more general analysis of weaker forms of score dependence based on copula models.

For the UNC–UIUC game, this approach yields the predictions illustrated in Figure 2. For each possible U on the horizontal axis we obtain a pair of scores whose vertical difference represents the margin of victory. For $\tau < 0.60$, UNC looks like it should be the clear winner, however, in the upper tail, that is in high scoring games, UIUC has a slight advantage. At the median, UNC has about a three point advantage, while at the first

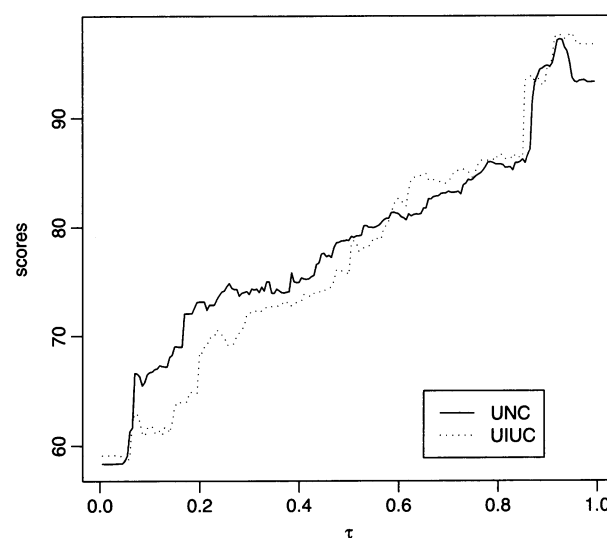


Figure 2. Estimated marginal quantile functions for the scores of the 2004–2005 NCAA final game between UNC and UIUC.

quartile their advantage is four, and at the third quartile UIUC has a slight edge.

It is tempting to attach some psychological or physiological interpretation to the values τ and U , but remembering that each score is the consequence of both an offensive rating and a defensive rating the model makes no such judgments. In the conventional paired comparison models based on mean performance there is a built-in assumption that the ability of teams differs by a constant factor and this difference applies over the whole range of the distribution. However, in the quantile regression version of the model, it is quite possible that one team can be more variable in their performance on offense, or on defense, or both, while another, more consistent team can be superior with high probability. This flexibility of the model raises some new questions for prediction. We do not want to make a prediction of the winner of the game based on only what is predicted to happen "at the median." Nor do we want to make a prediction about the point spread based on estimates of mean scores. For the UNC–UIUC game, there seems to be a clear signal to choose UNC if one is asked to pick a winner, but this may be too easy. What if we are asked to predict whether UNC will beat the posted Las Vegas point spread of two points? This question leads us toward a more realistic prediction that incorporates the dependence between scores.

We have considered, so far, two extreme models of scoring: one in which the scores are independent realizations from our marginal quantile functions, the other in which the two scores are deterministically linked. A more sensible view is the one provided by the copula model mentioned earlier. Given our estimated copula model, we can draw a pair of independent standard uniform random variables, U_i , U_j and evaluate,

$$(\hat{s}_i, \hat{s}_j) = (\hat{Q}_{ig}(U_i), \hat{Q}_{jg}(U_j)).$$

These uniforms, since they are generated from the estimated copula model, are dependent, and consequently, the generated scores are also dependent, but not comonotonic as in the situation illustrated in Figure 2. Repeated evaluations like this yield a predictive distribution for the scores of the game, from which we can make various predictions. For example, the estimated probability of team i beating team j by more than a specified point spread Δ is simply the proportion of generated points on the right side of the line $s_i - s_j = \Delta$.

In Figures 3 to 5 we illustrate the predictions of the model for 48 of the 64 games of the 2005 NCAA tournament based on estimation of the model using games up to and including the tournament round prior to the game. Eleven of the tournament games involved teams for which our season information is insufficient to estimate ratings; five additional first round games are dropped to reduce the plotting region. We follow the procedure described previously to simulate $G = 10,000$ realizations of the scores for each game from the estimated model. These scores are then projected on the $(-1, 1)$ axis to produce winning point margins for the G games and densities are estimated for each game using the default kernel method of R. This simulation method of producing predictive densities is closely related to the "rearrangement" methods for monotoneization of conditional quantile estimates introduced recently by Chernozhukov, Fernández-Val, and Galichon (2006). Vertical grey lines in these figures depict the zero reference value, black

lines indicate the actual score of the game, and the edge of the shaded region indicates the Las Vegas closing point spread announced for the game.

The first thing to say about these figures is that there are substantial differences in the dispersion and shape of the estimated densities as well as their location. Thus, the usual location shift hypothesis that underlies the conventional paired comparison models seems to require some reassessment. An examination of the position of the announced point spreads shows that they are usually "toward the mode" of the estimated densities and away from the grey "toss-up" line. Whether this should be interpreted as a vindication of our model and estimation method, or as the cleverness of the Las Vegas gambling establishment, we leave to the learned reader. The black lines designating the actual outcome of the games are considerably more dispersed, as they should be.

6.1 Bracketology and Tournament Survival

Of course predicting outcomes of the games played in the actual 2005 tournament as it evolved does not provide adequate guidance for filling in the tournament bracket *ex ante*. For this we will need predictions for pairings that might have occurred, but did not happen to occur in the 2005 tournament. No problem. Any pair of teams for which we have data can be pitted against one another with the scores generated according to the model and probabilities estimated. To this end, we simulate 1000 realizations of the 2005 tournament starting in every case from the actual seedings as announced on Selection Sunday. This exercise is slightly complicated by the fact that for some teams we do not have adequate season data to estimate ratings. However, since these teams were generally obscure and did not fare well in the first round, we assume that they will lose and their opponent was given a bye.

Simulating 1000 instances of March Madness 2005 takes about as long as a commercial time-out using a somewhat antiquated Mac G5. Given the outcomes of these simulated tournaments we can easily compute the number of successful rounds for each team in each replication; from this, survival curves can be estimated for each team. These survival curves are shown in Figure 6 where they are ordered by mean survival time. Recall that mean survival time can be expressed as the sum of the survival probabilities for the six rounds, and therefore, the area under the curves in the figure give a visual indication of the expected round that each team exits from the tournament. On this criterion UNC again comes first with mean 4.025, while UIUC is second with mean 3.905, Duke University is next with 2.953. These values provide only one of many possible ways to assess performance in the tournament. Another is to simply look at the probability that each team has of winning the tournament. On this criterion we have UNC with probability 0.318 and UIUC with 0.233; the next most likely winner is Duke at 0.083, and then we have the University of Louisville and Michigan State University (MSU) both with probabilities of 0.057.

7. IF YOU'RE SO SMART, WHY AREN'T YOU RICH?

We will be remiss if we fail to address *one last question* that looms large over any enterprise such as ours. In this concluding section, we explore several betting strategies based on the

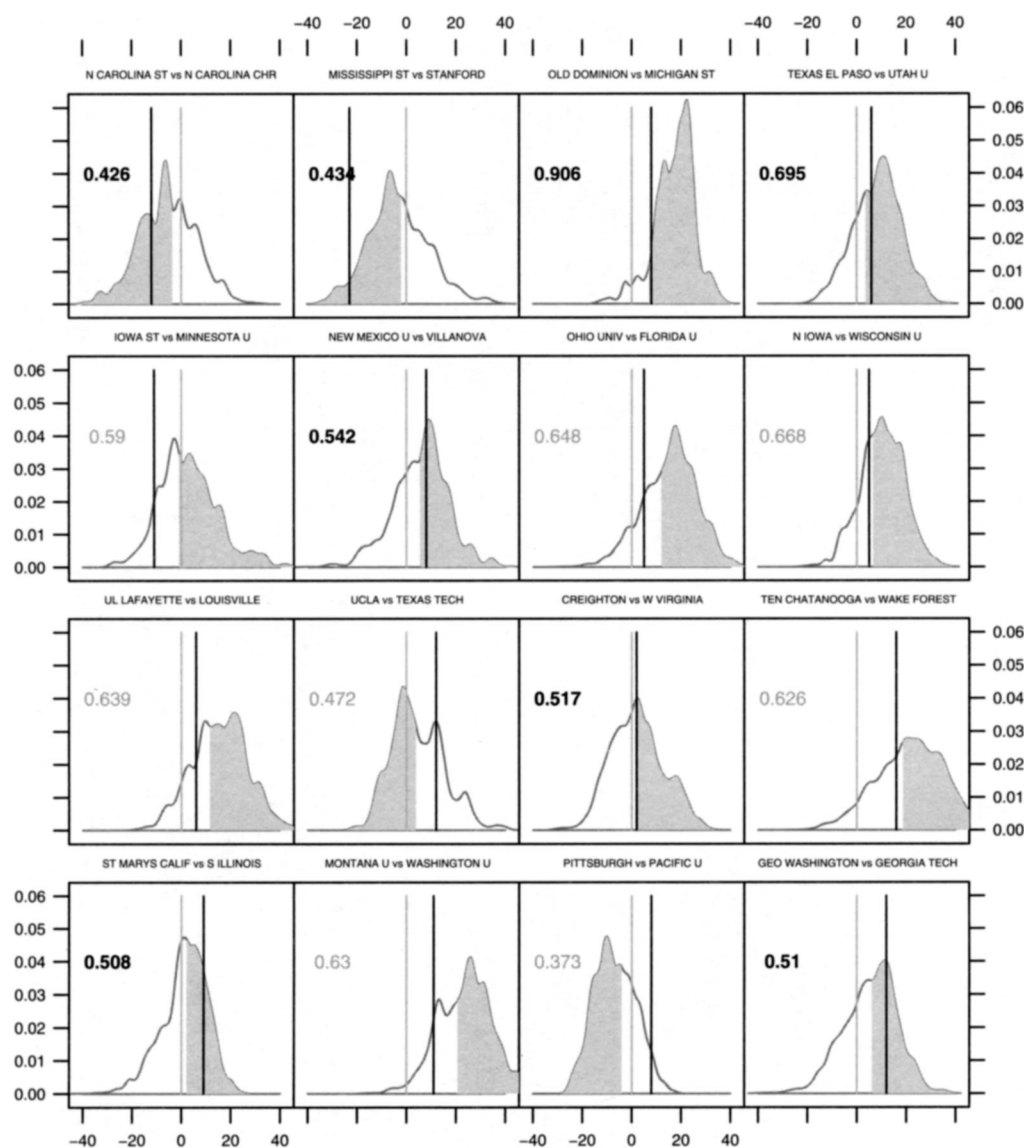


Figure 3. Predictive densities for the point spread of 2005 NCAA tournament games.

foregoing results and evaluate how they would have done based on the 2005 tournament. Predictions are based on updated estimates of the model including games of the previous round. Thus, first round game predictions use only the regular season data, second round games use this data plus the results of the first round games, and so forth.

7.1 Betting on the Point Spread

We begin by considering betting on point spreads. We can—as we already note following Equation (4.1)—estimate the probability of beating the announced point spread in any particular game. As long as this π_{ij} is different from $1/2$ there is a temptation to put money on the table. Of course we do not “know” the relevant p ’s for the NCAA games, but we have our estimates and may hope that strategies based on known p ’s might perform decently for estimated ones. In unfavorable games bold play is optimal, as we know from Dubins and Savage (1965), and this dictates placing one very large bet on the

game with the largest divergence between \hat{p} and $1/2$, but strategies for favorable games with uncertain probabilities are more complex. Breiman (1961) provided an elegant introduction to this subject.

Before going any further, we may want to check whether there is any merit in the conjecture that betting according to the model on the games of the 2005 NCAA tournament yields a profit. Returning to Figures 3 through 5, we explore this conjecture game by game: for each game we indicate the estimated π_{ij} for the *closing* point spread. This value corresponds to the area shaded in the figure. For $\pi_{ij} < 0.5$, we bet on the team *after* the “versus” in the panel title, since it indicates that there is a better than 50–50 chance that this team will beat the point spread. Thus, if the black vertical line denoting the winning margin of the actual game falls in the shaded region our bet will be successful, otherwise it will not be successful.

For example, for the game between the University of Wisconsin and UNC in the upper left corner of Figure 6, the point spread was 11, the model predicts that the probability of UNC winning by 11 or more is only 0.490 so we bet on Wisconsin,

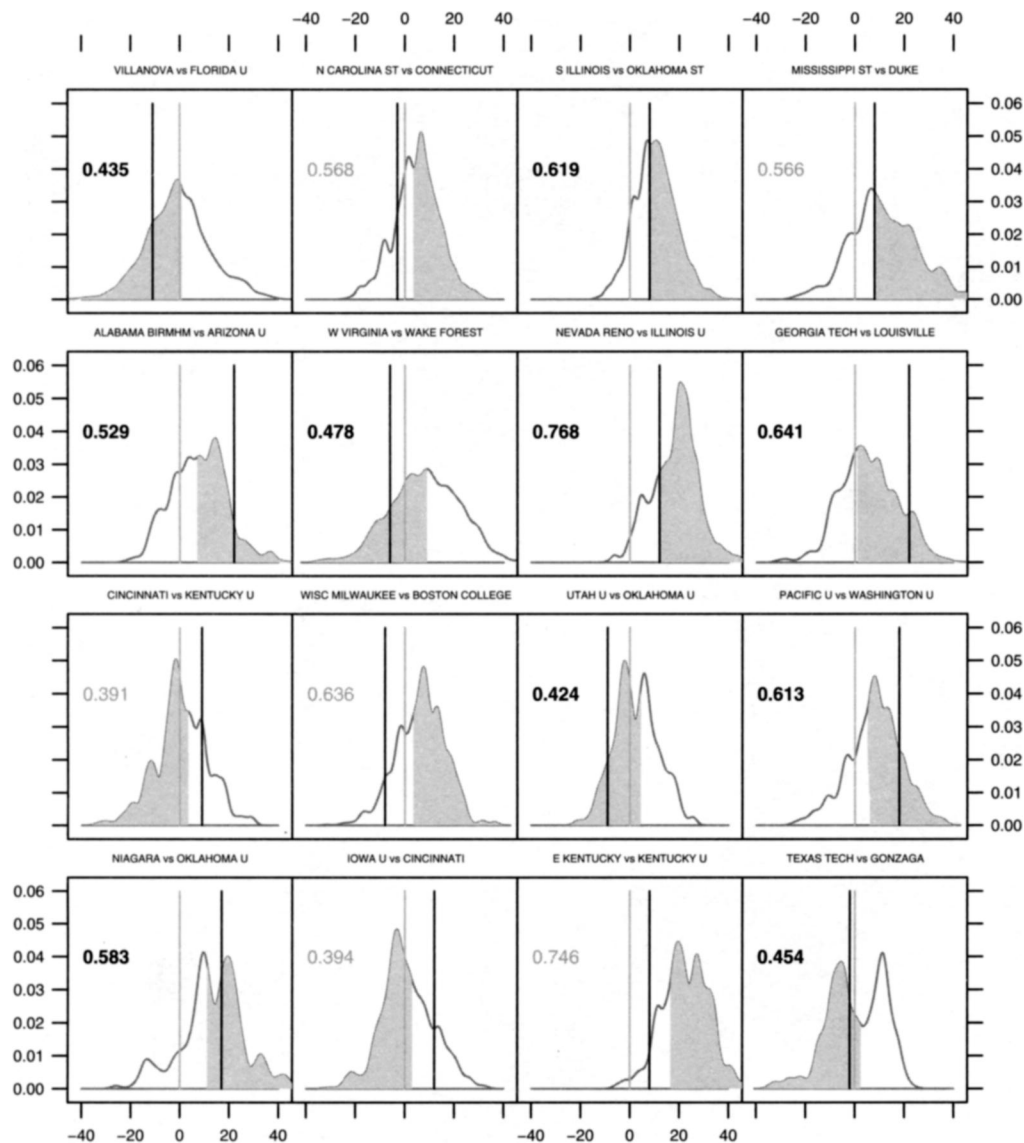


Figure 4. Predictive densities for the point spread of 2005 NCAA tournament games.

and since UNC only won by six points, we collect. With this visual heuristic in place, we can scan through the panels of Figures 3 to 5. As an aid to this scanning we indicate the probabilities appearing in each panel with the successful predictions in bold black and the unsuccessful in grey. What we find is mildly encouraging: in 28 out of 48 games we select the winning side of the point spread. Early games have a less impressive performance with only 8 of the first 16 games in Figure 3. But, of the final 32 games, we have 20 successful predictions, a finding that may simply reflect the ancient and canonical answer to our canonical question: "It is better to be lucky, than smart."

In fact, closer examination reveals that there is one game, West Virginia versus Louisville, that the posted point spread was predicted precisely. Such "push" games are treated, by convention, as if there was no bet, so money is refunded. Thus, we should properly consider the model to predict 27 out of 47, giving a frequency of success of 0.574. This is certainly not significantly different from 0.5 at conventional levels of significance. The p -value of an exact test is only 0.38, but given that it costs

\$110 to place a \$100 bet, our 0.574 frequency implies that we will have an expected gain of about \$10.50 on each \$100 bet.

We might want to ask whether this observed frequency of success is consistent in some way with the predicted frequencies of the model. In repeated trials, the model purports to predict that the frequency of success will be $\hat{\pi}_{ij}$ for the games between teams i and j , so the mean of the $\hat{\pi}_{ij}$'s provides a reference level for our empirical frequency of success. This mean is 0.603 and is not terribly different from the observed frequency.

A natural question at this point is: "OK, you seem to be doing somewhat better than coin flipping, but 48 games is a small sample and you have not shown that you cannot do just as well with the classical Gaussian model and least-squares estimation." The densities that appear in Figures 3 and 4 are all quite unimodal and roughly symmetric looking, so maybe all this flexibility of the quantile regression model is just contributing to noisier estimation of the ratings effects. It is easy to evaluate these claims. We simply estimate a least-squares version of the rating model, along with the covariance matrix for the scoring pairs. This yields a Pearson correlation of 0.44 for the scores

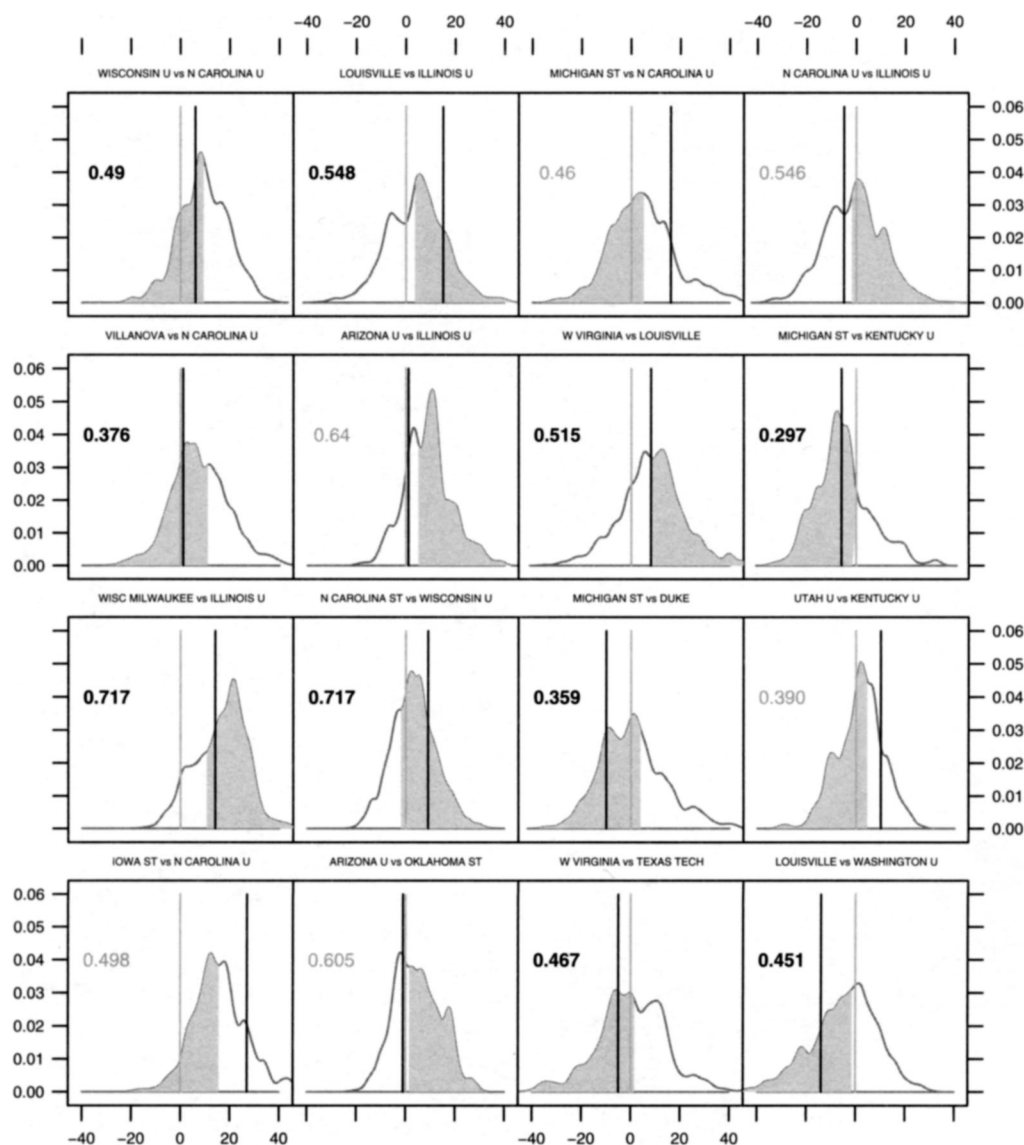


Figure 5. Predictive densities for the point spread of 2005 NCAA tournament games.

and a standard deviation of the score difference of 9.93 points. Given these estimates we can estimate probabilities of the actual score difference exceeding the closing point spread under Gaussian conditions. These estimated probabilities are plotted against the corresponding estimates from the quantile regression model in Figure 7. Not surprisingly, they agree quite well, but for eight games appearing in Quadrants II and IV in the scatterplot, the two models disagree on what side of the point spread one should bet on. These games split four and four so the least-squares model also has 27 of 47 successes. We should emphasize that, despite the similar overall performance of the two sets of predictions, the underlying models are very different and the predictions actually conflict in roughly one out of six games. The added flexibility of the quantile regression model is likely to incur some cost of increased variability, and as usual, performance is determined by a balance of bias and variance considerations. As sample size increases, bias inevitably dominates and we anticipate that the proposed quantile regression approach will dominate.

The next natural question is: What accounts for this modest violation of the Hayek hypothesis? Granted, gamblers may not be completely *au fait* regarding quantile regression, but surely they are familiar with the fundamentals of least squares? Where are they going wrong? One contributing factor was suggested by Camerer (1989) who argued that the market tends to favor teams with strings of wins in the apparent belief that teams get “hot”—while the evidence suggests, on the contrary, that such dependence is illusory. Our modeling, since it assumes independent realizations, may, therefore, have some advantage by avoiding this momentum misapprehension. But there are, no doubt, many other contributing factors.

7.2 Betting on the Over/Under

Aficionados will be aware that one can bet not only on scores differences, but also on their sum. Like the point spread, a “total” is posted by bookies and one can bet that the sum of the two scores of a given game will exceed or fall short of this number. This is the so-called over/under. While betting on point spreads



Figure 6. Survival functions for the 2005 NCAA tournament. Obtained by simulation of the quantile regression rating model and using the Frank copula model to generate random uniforms.

has an inherent element of partisanship, one might imagine that betting on totals will be an act of pure rationality unsullied by the emotions of geography, and thus less likely to reveal market inefficiencies.

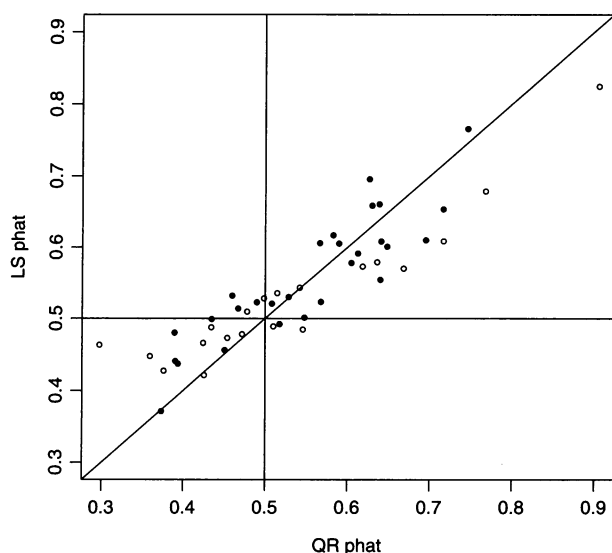


Figure 7. Estimated probabilities of success against the point spread.

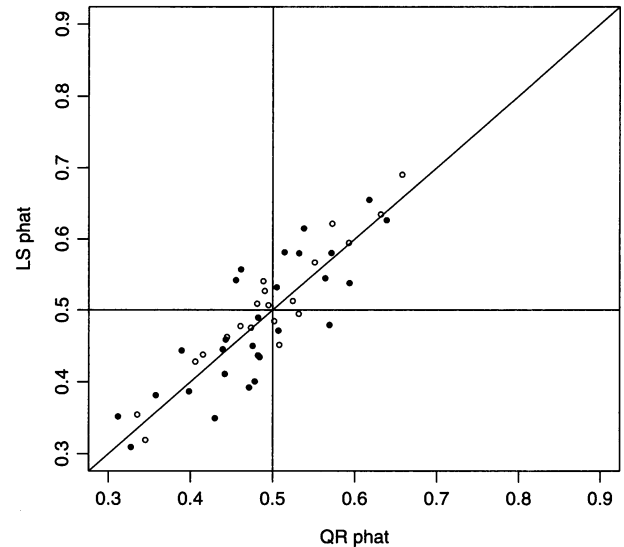


Figure 8. Estimated probabilities of success on the over/under.

As an additional test of the merit of the model, we estimate densities for the “totals” for the 48 games of the NCAA tournament based on the qualitative reasoning (QR) model. These figures are omitted due to space constraints, but are available in the electronic Appendix to the article, together with some graphical and numerical results on overall ratings of the 2005 tournament teams. Two games are ambiguous. One is the West Virginia versus Wake Forest game for which the posted total was 152.5 and which West Virginia won 111 to 105; the model predicted correctly. In the MSU–UNC semifinal game the total was 158 and the final score was 71 to 87; in such “push” cases the original stake is refunded. Of the 47 remaining games under consideration, 27 were correctly predicted by the model. Curiously the model’s prediction of its success rate is almost the same with the mean of the $\max\{\hat{\pi}_{ij}, 1 - \hat{\pi}_{ij}\}$ only 0.574.

Returning to the question of whether least-squares estimates can deliver a similar performance, we re-estimated probabilities of success on over/under bets using the mean model. Figure 8 displays the scatterplot of the predicted probabilities from the two models, again with solid points indicating successes and open points indicating failures of the QR model. As with point spreads, the two models produce quite similar estimated probabilities. There is a conflict over what side to bet on in 11 of the 48 games, and of these the outcomes split four and seven, so the least-squares version of the model predicts 30 of 47 correctly. This is astonishingly lucky, but probably should not be taken as further evidence, as if any were needed, that Gauss was smart.

7.3 Betting on Simple Parlays

Combination bets on both point spreads and totals are called parlays, and to conclude this section we evaluate a simple example of such a betting strategy. Again, we employ the model to predict which of the four possible pairs of outcomes, over/under the point spread, over/under the total, is most likely to occur. If the point spreads and the totals were fiendishly well set to reflect the objective probabilities we expect that each of the four quadrants is assigned probability 0.25. But bookies could care less about “objective probabilities”—as long as the money is

balanced on their bets—they collect the vigorish, typically arising from the fact that the bettor puts up \$110 to bet \$100. In our 48 games of the NCAA tournament, the mean of the maximal quadrant probabilities is 0.364. Betting on these parlays wins in only 13 out of the 48 games, or in 0.27% of the cases, better than guessing randomly, but just barely.

To evaluate a comparable strategy using the least-squares estimates we first evaluate the predicted mean spread and total according to the estimated model then, using the estimated covariance matrix for the scores, we compute the orthant probabilities of the posted point spread and total given the bivariate normal model with this mean and covariance matrix. Again, we choose the largest of these four probabilities and bet on this quadrant. The QR and least-squares (LS) models agree on which quadrant to bet on in only 10 of the 48 games, but the least-squares bets get 17 out of 48 games right, for a rather impressive 0.35 success rate.

8. REFINEMENTS

As we suggest in the Introduction, there are many potential refinements of the methods we introduce. The rather profligate parameterization of the model will undoubtedly benefit from some judicious form of regularization—designed to shrink toward common ratings, or toward prior season ratings, for example. It may also prove useful to reweight the estimation of the rating model to give more credence to games toward the end of the season. Much more can be said about gambling strategies in this context, but these topics will be left as grist for future grinding.

9. CONCLUSIONS

A more flexible variant of the classical paired comparison model of mean ratings is described and evaluated based on NCAA college basketball data. The model permits a wide variety of heterogeneity in teams' offensive and defensive "ability" and provides a simple mechanism for making predictions about the subsequent performance of the teams. The model is estimated on a sample 2940 regular season games involving 232 teams. Out-of-sample predictive performance of the model is evaluated based on 48 games of the 2005 NCAA Tournament. This evaluation reveals mildly favorable betting opportunities against posted point spreads and scoring totals for these games. Predictions based on a comparable mean rating model estimated by least squares have a somewhat better performance. In defense of the added complexity of the quantile regression form of the paired comparison model we offer the *cri de coeur* of every sports fan: "Wait until next year!"

SUPPLEMENTAL MATERIALS

Appendix: An Appendix containing additional predictive densities for game point totals and team ratings is available from the *JBES* supplementary materials website. (Eapp.pdf)

ACKNOWLEDGMENTS

This research was partially supported by NSF grant SES-05-44673. The authors thank and express their appreciation to Dan Bernhardt and Steve Heston for advice and for providing the data analyzed in Section 6, to Paul Murrell for his R grid graphics package (Murrell 2006), to Jun Yan for his R copula package (Yan 2007), and to Xuming He, Steve Portnoy, Victor Chernozhukov, and two anonymous referees for helpful comments.

[Received April 2007. Revised April 2008.]

REFERENCES

- Bassett, G. (1997), "Robust Sports Ratings Based on Least Absolute Values," *The American Statistician*, 51, 99–105. [27]
- Breiman, L. (1961), "Optimal Gambling Systems for Favorable Games," *Proceedings of the Fourth Berkeley Symposium on Probability and Mathematical Statistics*, 1, 65–78. [31]
- Camerer, C. F. (1989), "Does the Basketball Market Believe in the 'Hot Hand'?" *American Economic Review*, 79, 1257–1261. [33]
- Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2006), "Quantile and Probability Curves Without Crossing," preprint. [30]
- David, H. A. (1988), *The Method of Paired Comparisons* (2nd ed.), London: Griffin. [27]
- Dubins, L. E., and Savage, L. J. (1965), *How to Gamble if You Must: Inequalities for Stochastic Processes*, New York: McGraw-Hill. [31]
- Gutenbrunner, C., and Jurečková, J. (1992), "Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics," *The Annals of Statistics*, 20, 305–330. [29]
- Hayek, F. A. (1945), "The Use of Knowledge in Society," *American Economic Review*, 35, 519–530. [26]
- Koenker, R. (2005), *Quantile Regression*, London: Cambridge University Press. [28,29]
- (2006), "quantreg: An R Package for Quantile Regression," available at <http://www.r-project.org>. [28]
- Koenker, R., and Ng, P. (2005), "A Frisch–Newton Algorithm for Sparse Quantile Regression," *Acta Mathematicae Applicatae Sinica*, 21, 225–236. [28]
- (2006), "SparseM: An R Package for Sparse Linear Algebra," available at <http://www.r-project.org>. [28]
- Murrell, P. (2006), *R Graphics*, London: Chapman & Hall/CRC. [35]
- Portnoy, S. (1991), "Asymptotic Behavior of Regression Quantiles in Non-Stationary, Dependent Cases," *Journal of Multivariate Analysis*, 38, 100–113. [28]
- Portnoy, S., and Koenker, R. (1997), "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error versus Absolute-Error Estimators" (with discussion), *Statistical Science*, 12, 279–300. [28]
- Sauer, R. D. (1998), "Economics of Wagering Markets," *Journal of Economic Literature*, 36, 2021–2064. [26]
- Yan, J. (2007), "copula: An R Package for Multivariate Dependence With Copulas," available at <http://www.r-project.org>. [35]