The Home-Court Advantage: How Large Is It, and Does It Vary From Team to Team?

David A. HARVILLE and Michael H. SMITH

College basketball fans often speculate on the extent and nature of the home-court advantage. In this article, it is shown how questions about the size of the home-court advantage and about team-to-team differences in the home-court advantage can be addressed by fitting three different linear models to the outcomes of college basketball games. This approach was applied to data from the 1991–1992 college basketball season. The advantage of playing at home (in relation to playing on a neutral court) was estimated to be 4.68 ± 0.28 points. The effect of team-to-team differences in the home-court advantage was found to be relatively small.

KEY WORDS: Analysis of variance; College basketball: Least squares; Prediction; Ranking.

1. INTRODUCTION

College basketball has become one of America's most popular spectator sports. Of the colleges and universities that belong to the National Collegiate Athletic Association (NCAA), approximately 300 field teams that compete in Division I (which is the highest level of competition). Few followers of the sport can resist the temptation to make predictions about the outcomes of the games; in many cases, these predictions are translated into wagers, with stakes that range from "bragging rights" to large sums of money.

Various individuals or organizations periodically compile ratings or rankings of the teams. Some of the ratings or rankings are widely publicized and often generate nearly as much interest as the games themselves. What may be the most important of all the ratings, however, are not made available to the public. These are ratings compiled by the NCAA itself, on the basis of something called the rating percentage index (RPI). They are provided to the NCAA Division I Men's Basketball Committee to assist the committee in completing and seeding the 64-team field for the postseason Division I men's basketball tournament. A college or university's visibility and financial well being can be directly or indirectly affected by its team's inclusion (and success or failure) in the tournament.

In predicting the outcomes of college basketball games, it is thought to be very important to account for the home-court advantage (as well as the relative abilities of the opposing teams). The home-court advantage refers to the net effect of several factors that may have an (generally positive) effect on the play of the home team and an (generally negative) effect on the play of the visiting team. These factors include (1) crowd support, which may affect the teams indirectly (e.g., through its effect on the referees) as well as directly; (2) distinctive features of the playing facilities (e.g., with regard to the baskets, basket supports, and backboards, the composition of the playing surface, and the shooting background), to which the home team is typically more accustomed than the visiting team; and (3) travel, which can have a disruptive effect on the sleeping and eating patterns of the visiting players.

It is generally considered less important to account for the home-court advantage in rating or ranking teams (than in predicting the outcomes of games). In fact, the RPI makes no allowance whatsoever for the home-court advantage. For a team that plays nearly the same number of games "at home" as "on the road," the effects of the home-court advantage may (over the course of the season) average out. There are many teams, however, whose schedules are very unbalanced with regard to relative numbers of home and away games. In particular, teams whose home games tend to be poorly attended or whose television rights are not especially valuable may be forced for financial reasons to play more away games than those teams that have large followings. To ignore this imbalance in ranking or rating teams is to favor implicitly (and unfairly) those teams that have a preponderance of home games.

It is often maintained that some teams have significantly greater home-court advantages than others. Some reasons why a team might have an exceptionally large home-court advantage include (1) its home games might be played at a high-altitude location (e.g., University of Colorado), (2) the fans attending its home games might be unusually numerous and/or exceptionally boisterous and enthusiastic, and (3) its players might be inexperienced (in which case the presence or absence of crowd support might have an exaggerated effect).

How large is the home-court advantage? Has the tendency for the teams with the largest followings (which are typically the most proficient teams) to play more games at home than away created an exaggerated impression of the size of the home-court advantage? Are there differences in the home-court advantage from team to team? And if such differences exist, what are their magnitudes, and do the better teams tend to have the greater home-court advantages?

In what follows, we address these questions by fitting various linear models to the outcomes of games played during the 1991–1992 college basketball season. In Sections 2 and 3, we describe the data and introduce the models that were fitted to these data. In Section 4, we present the results of various least squares analyses and use them to make some inferences about the home-court advantage. Finally, in Section 5, we discuss some modifications and refinements to the fitted models.

David A. Harville is Professor, Department of Statistics, Iowa State University, Ames, IA 50011. Michael H. Smith is Statistician, Hoffmann-LaRoche, Inc., 340 Kingsland Street, Nutley, NJ 07110. The authors are indebted to two reviewers for their helpful comments.

As a teacher and student, we have found the application of linear models to basketball scores to be very instructive and recommend the use of this application in graduate-level courses in linear models. College basketball gives rise to several questions that can be addressed by linear statistical inference and that may generate interest among the students. And many students are sufficiently knowledgeable about college basketball that they can make informed judgments about the choice of model or models. Moreover, the models encountered in this application have some nonstandard features and can be very helpful in illustrating such concepts as estimability, connectedness, dummy variables, constraints on parameters or least squares solutions, parsimony, simultaneous inference, practical versus statistical significance, and inference about the expected value of a future observation versus (predictive) inference about the observation itself.

2. DATA

The data for the study were accumulated during the course of the 1991–1992 college basketball season by recording the outcomes of the various games as they were reported in newspapers. To facilitate this process, 155 (Division I) teams were selected, and the data set was restricted to those games in which both opponents were among the selected teams.

The selected teams included the (144) teams from the following 16 conferences: Atlantic Coast, Atlantic 10, Big East, Big Eight, Big Ten, Big West, Great Midwest, Metropolitan, Midwestern Collegiate, Mid-American, Missouri Valley, Pacific-10, Southeastern, Southwest, Sun Belt, and Western Athletic. The other 11 of the 155 selected teams consisted of the following independents (teams not affiliated with a conference): California State–Northridge, California State–Sacramento, Chicago State, Missouri–Kansas City, North Carolina–Greensboro, Northeastern Illinois, Notre Dame, Pennsylvania State, Southern Utah, Wisconsin–Milwaukee, and Youngstown State. These 16 conferences and 11 independent teams were chosen on the basis of prominence and on the interconnectedness of their schedules.

The study was restricted to games played during the regular season (i.e., games played during the 64-team Division I championship tournament were not included). A total of 1,678 games qualified for inclusion. For each game, the publication *Official 1992 NCAA Basketball* (NCAA 1991) was used to determine which (if either) team was the home team or whether the game was played on a neutral court.

3. STATISTICAL FORMULATION

Consider the problem of making statistical inferences about the home-court advantage, based on data (like those described in Section 2) consisting of the outcomes of n college basketball games played among t teams during the course of a single season. The questions raised (in Section 1) about the home-court advantage can be formulated in terms of various linear models that might be applied to the n differences in score. (The modeling of the data is greatly facilitated by restricting attention to statistical procedures that depend on the data through the differences in score only.)

Let r_{ij} represent the number of games in which the ith team is the home team and the jth team is its opponent, and let y_{ijk} represent the difference in score between the ith and jth teams in the kth of these r_{ij} games. For any game played on a neutral court, we arbitrarily (for notational purposes only) label one of the two opposing teams the home team. Denote by x_{ijk} a dummy variable that equals 0 if the ijkth game is played on a neutral court, and equals 1 otherwise (i.e., if the ith team is truly the home team).

Consider the following three alternative models:

$$\text{Model 1: } y_{ijk} = \beta_i - \beta_j + e_{ijk};$$

$$\begin{aligned}\text{Model 2: } y_{ijk} &= \lambda + \beta_i - \beta_j + e_{ijk} \quad &\text{if } x_{ijk} = 1\\ &= \beta_i - \beta_j + e_{ijk} &\text{if } x_{ijk} = 0;\end{aligned}$$

$$\begin{aligned}\text{Model 3: } y_{ijk} &= \alpha_i - \beta_j + e_{ijk} \quad &\text{if } x_{ijk} = 1\\ &= \beta_i - \beta_j + e_{ijk} &\text{if } x_{ijk} = 0.\end{aligned}$$

Here, $\alpha_1, \ldots, \alpha_t, \beta_1, \ldots, \beta_t$, and λ are unknown parameters, and the e_{ijk}'s are uncorrelated random residual effects having mean 0 and common unknown positive variance σ^2.

Stefani (1977) used Model 1 as a basis for rating college basketball teams and college and professional football teams (and for predicting the outcomes of basketball and football games). Model 1 makes no allowance for a home-court advantage.

Stefani (1980) used Model 2 to improve on his earlier work, and Harville (1977, 1978) used a variation on Model 2 (in which $\beta_1, \ldots, \beta_t$ were regarded as random variables rather than as unknown parameters) in devising a rating and prediction system. Model 2 allows for the possibility of a home-court advantage. It is implicit in Model 2 that the home-court advantage does not vary from team to team and that the expected difference in score between any two teams in a game played on a neutral court is halfway between that in a game played on the first team's home court and that in a game played on the second team's home court.

Like Model 2, Model 3 provides for the existence of a home-court advantage; unlike Model 2, however, it allows for the possibility that the home-court advantage may vary from team to team. It is implicit in Model 3 that the expected difference in score between any two teams in a game played on a neutral court equals the difference in the expected differences in score in games played by them against a common opponent on the opponent's home court.

Let $\mathbf{y}$, $\mathbf{e}$, and $\mathbf{x}$ represent the $n \times 1$ vectors with ijkth elements y_{ijk}, e_{ijk}, and x_{ijk}, respectively. Moreover, for $m = 1, \ldots, t$, let $\boldsymbol{u}_m$ represent the $n \times 1$ vector whose ijkth element equals 1 if $m = i$ and $x_{ijk} = 1$ and equals 0 otherwise; let $\boldsymbol{v}_m$ represent the $n \times 1$ vector whose ijkth element equals 1 if $m = i$ and $x_{ijk} = 0$, equals -1 if $m = j$, and equals 0 otherwise; and let $\boldsymbol{w}_m = \boldsymbol{u}_m + \boldsymbol{v}_m$ represent the $n \times 1$ vector whose ijkth element equals 1 if $m = i$, equals -1 if $m = j$, and equals 0 otherwise. Then, Models 1–3 can be rewritten in matrix form as $\mathbf{y} = \boldsymbol{X}_i\boldsymbol{\beta}^{(i)} + \mathbf{e}$ $(i = 1, 2, 3)$, respectively, where $\boldsymbol{X}_1 = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t)$, $\boldsymbol{\beta}^{(1)} = (\beta_1, \ldots, \beta_t)'$, $\boldsymbol{X}_2 = (\mathbf{x}, \boldsymbol{X}_1) = (\mathbf{x}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_t)$, $\boldsymbol{\beta}^{(2)} = (\lambda, \beta_1, \ldots, \beta_t)'$, $\boldsymbol{X}_3 = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_t, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_t)$ and $\boldsymbol{\beta}^{(3)} = (\alpha_1, \ldots, \alpha_t, \beta_1, \ldots, \beta_t)'$.

For illustration, consider a small example that encompasses six games and three teams, in which Team 1 plays twice at home, losing to Team 2 by 1 point and defeating Team 3 by 9 points; Team 2 plays twice at home, defeating Team 1 by 15 points and Team 3 by 20 points, and once on a neutral court, defeating Team 3 by 6 points; and Team 3 plays once at home, losing to Team 2 by 7 points. In this example, $n = 6$, $t = 3$, $r_{31} = 0$, $r_{12} = r_{13} = r_{21} = r_{32} = 1$, $r_{23} = 2$:

$$\mathbf{y} = \begin{pmatrix} y_{121} \\ y_{131} \\ y_{211} \\ y_{231} \\ y_{232} \\ y_{321} \end{pmatrix} = \begin{pmatrix} -1 \\ 9 \\ 15 \\ 20 \\ 6 \\ -7 \end{pmatrix}$$

$$\boldsymbol{X}_1 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

$$\boldsymbol{X}_2 = \begin{pmatrix} 1 & 1 & -1 & 0 \\ 1 & 1 & 0 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 1 \end{pmatrix}$$

$$\boldsymbol{X}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix}$$

where, in the game played by Teams 2 and 3 on a neutral court, Team 2 is designated the home team.

Clearly,

$$\sum_{m=1}^{t} \boldsymbol{w}_m = \sum_{m=1}^{t} \boldsymbol{u}_m + \sum_{m=1}^{t} \boldsymbol{v}_m = 0, \qquad (1)$$

implying in particular that none of the three model matrices $\boldsymbol{X}_1, \boldsymbol{X}_2$ and $\boldsymbol{X}_3$ is of full column rank. Subsequently, we assume that when the n games are regarded as forming a two-way crossed classification (by home team and visiting team), the data are connected and would be connected even if those games played on a neutral court were disregarded—the data described in Section 2 satisfy this assumption (as do the data in the illustrative example). Then, as a consequence of standard results on connectedness (e.g., Searle 1971, sec. 7.4), we have that

$$\begin{aligned} \operatorname{rank}(\boldsymbol{X}_1) &= t - 1, \\ \operatorname{rank}(\boldsymbol{X}_2) &= t, \\ \operatorname{rank}(\boldsymbol{X}_3) &= 2t - 1. \end{aligned} \qquad (2)$$

It is evident from Equation (1) that $\boldsymbol{X}_1 \mathbf{1} = \mathbf{0}$ (where $\mathbf{1}$ represents a column vector of ones) and from Equation (2) that the number of columns in $\boldsymbol{X}_1$ exceeds rank $(\boldsymbol{X}_1)$ by just one. Thus, it follows from well-known results on estimability that a linear combination $\sum_{m=1}^{t} d_m \beta_m$ of the parameters in Model 1 is estimable if and only if $(d_1, \ldots, d_t)\mathbf{1} = 0$ or equivalently if and only if $\sum_{m=1}^{t} d_m = 0$. A similar line of reasoning leads to the conclusion that a linear combination $c\lambda + \sum_{m=1}^{t} d_m \beta_m$ of the parameters in Model 2 is estimable if and only if $\sum_{m=1}^{t} d_m = 0$, and a linear combination $\sum_{m=1}^{t} c_m \alpha_m + \sum_{m=1}^{t} d_m \beta_m$ of the parameters in Model 3 is estimable if and only if $\sum_{m=1}^{t} c_m + \sum_{m=1}^{t} d_m = 0$ or equivalently if and only if $\sum_{m=1}^{t} c_m = -\sum_{m=1}^{t} d_m$. Note that included among the estimable linear combinations of the parameters in Model 3 is any contrast in $\alpha_1, \ldots, \alpha_t$ and any contrast in $\beta_1, \ldots, \beta_t$ (i.e., any linear combination $\sum_{m=1}^{t} c_m \alpha_m$ of $\alpha_1, \ldots, \alpha_t$ such that $\sum_{m=1}^{t} c_m = 0$ and any linear combination $\sum_{m=1}^{t} d_m \beta_m$ of $\beta_1, \ldots, \beta_t$ such that $\sum_{m=1}^{t} d_m = 0$).

The home-court advantage can be formally defined as the expected difference in score in a game played by a team, say the ith team, on its home court minus the expected difference in score in a game played by the ith team on a neutral court (against the same opponent). In Model 1, this quantity equals zero; in Model 2, it is represented by the estimable parameter λ; and in Model 3, it is represented by the estimable function $\alpha_i - \beta_i$. Note that the estimable function $(1/t) \sum_{i=1}^{t} (\alpha_i - \beta_i)$ (in Model 3) represents the average (over the t teams) home-court advantage.

Under Model 2, the null hypothesis that there is no home-court advantage can be formulated as the null hypothesis $H_0^{(1)} : \lambda = 0$, and under Model 3, it can be formulated as the null hypothesis $H_0^{(2)} : \alpha_1 - \beta_1 = \cdots = \alpha_t - \beta_t = 0$. Under Model 2, the alternative hypothesis might be that $\lambda \neq 0$ or alternatively that $\lambda > 0$; under Model 3, the alternative hypothesis might be that $\alpha_i - \beta_i \neq 0$ for at least one i or alternatively that $\alpha_1 - \beta_1 \geq 0, \ldots, \alpha_t - \beta_t \geq 0$, with $\alpha_i - \beta_i > 0$ for at least one i. Which alternative hypothesis is appropriate would depend on whether it were assumed that the home-court advantage is truly an advantage. The reduced model (e.g., Searle 1971, secs. 3.6 and 5.5) for either of the two null hypotheses $H_0^{(1)}$ or $H_0^{(2)}$ is Model 1.

The null hypothesis that the home-court advantage is the same for all t teams can be formulated (under Model 3) as $H_0^{(3)} : \alpha_1 - \beta_1 = \cdots = \alpha_t - \beta_t$ (which is less stringent than $H_0^{(2)}$). The reduced model for the null hypothesis $H_0^{(3)}$ is Model 2, as is evident on reexpressing α_i as $\beta_i + (\alpha_i - \beta_i)$ in Model 3.

There are various estimable functions of the parameters of Model 3 that reflect the expected performance level of a team, say the ith team, including $\beta_i - (1/t) \sum_{j=1}^{t} \beta_j$ (expected performance level as a visiting team or on a neutral court in relation to the average), $\alpha_i - (1/t) \sum_{j=1}^{t} \alpha_j$ (expected performance level as a home team in relation to the average), $\beta_i - (1/t) \sum_{j=1}^{t} \alpha_j$ (expected difference in score when playing on the road against an average home team), $\alpha_i - (1/t) \sum_{j=1}^{t} \beta_j$ (expected difference in score when playing at home against an average visiting team), and $\frac{1}{2}[\alpha_i + \beta_i - (1/t) \sum_{j=1}^{t} (\alpha_j + \beta_j)]$ (expected overall performance level in relation to the average). Note that

Table 1. Analysis of Variance

Source	*Degrees of freedom*	*Sum of Squares*	*Mean square*
Model 1	$t - 1 = 154$	$S_1 = 186,392.3$	
Model 3 $\mid$ Model 1	$t = 155$	$S_{3\mid 1} = 51,364.8$	$M_{3\mid 1} = 331.4$
Model 2 $\mid$ Model 1	1	$S_{2\mid 1} = 32,105.4$	$M_{2\mid 1} = 32,105.4$
Model 3 $\mid$ Model 2	$t - 1 = 154$	$S_{3\mid 2} = 19,259.4$	$M_{3\mid 2} = 125.1$
Residual	$n - 2t + 1 = 1369$	$S_e = 154,384.9$	$M_e = 112.8$
Total	$n = 1678$	$S_t = 392,142.0$	

NOTE: The symbol $S_{j|i}$ represents the difference between the residual sums of squares obtained by fitting Models i and j, respectively.

$\beta_i - (1/t)\sum_{j=1}^{t} \beta_j$ is an estimable function of the parameters of Models 1 and 2 as well as those of Model 3.

4. STATISTICAL INFERENCES

Statistical procedures for testing the three null hypotheses $H_0^{(1)}, H_0^{(2)}$, and $H_0^{(3)}$ and for making other kinds of inferences about the home-court advantage can be devised by using various results from the theory of linear models. In this section, we describe some such procedures and apply them to our data (i.e., to the data described in Section 2). Various computations relevant to testing $H_0^{(1)}, H_0^{(2)}$ and $H_0^{(3)}$ are summarized in Table 1 (both in general terms and as applied to our data).

Let us initially adopt Model 2 (and for purposes of devising hypothesis tests, suppose that the e_{ijk}'s are normally distributed). Then, an appropriate test of $H_0^{(1)}$ (vs. the two-sided alternative that $\lambda \neq 0$) is to reject $H_0^{(1)}$ for sufficiently large values of the statistic $F^{(1)} = M_{2|1}/M_p$, where $M_p = (S_{3|2} + S_e)/(n - t)$. Under the null hypothesis, $F^{(1)}$ has a (central) F distribution with 1 and $n - t$ df.

For our data, $n - t = 1523$, $M_p = 114.0$, $F^{(1)} = 281.59$, and the p value of the test is essentially zero. Furthermore, the least squares estimate of the home-court advantage λ is 4.68, with an estimated standard error of 0.28 (on 1523 df).

Let us now switch (from Model 2) to Model 3 (and for purposes of devising hypothesis tests, we continue to suppose that the e_{ijk}'s are normally distributed). Then, an appropriate test of $H_0^{(2)}$ (vs. the alternative that $\alpha_i - \beta_i \neq 0$ for at least one i) is to reject $H_0^{(2)}$ for sufficiently large values of the statistic $F^{(2)} = M_{3|1}/M_e$, and an appropriate test of $H_0^{(3)}$ is to reject $H_0^{(3)}$ for sufficiently large values of the statistic $F^{(3)} = M_{3|2}/M_e$. Under $H_0^{(2)}, F^{(2)}$ has a (central) F distribution with t and $n - 2t + 1$ df, and under $H_0^{(3)}, F^{(3)}$ has an F distribution with $t - 1$ and $n - 2t + 1$ df.

For our data, $F^{(2)} = 2.94$ and $F^{(3)} = 1.11$, and the p values of the tests of $H_0^{(2)}$ and $H_0^{(3)}$ are zero (for all practical purposes) and .183, respectively.

The overall home-court advantage is represented by λ in Model 2 and by $(1/t)\sum_{i=1}^{t}(\alpha_i - \beta_i)$ in Model 3. It seems clear from the results obtained under Model 2 (test of $H_0^{(1)}$, point estimate of λ, and estimated standard error of the point estimate) that not only is the overall home-court advantage different from zero, but it is large enough to be of practical importance. Essentially the same conclusion could be reached on the basis of Model 3 by testing the null hypothesis that the parametric function $(1/t)\sum_{i=1}^{t}(\alpha_i - \beta_i)$ equals zero and by obtaining the least squares estimate of this parametric function (along with its estimated standard error).

The observed value (1.11) of $F^{(3)}$ is not significant at standard reference levels like .01 or .05. Nevertheless, this value is somewhat higher than the expected value (approximately 1) of $F^{(3)}$ under the null hypothesis $H_0^{(3)}$, and consequently it can be regarded as providing some evidence of team-to-team differences in the home-court advantage.

Are the team-to-team differences (if any) in the home-court advantage large enough to be of practical importance? To answer this question, it may be helpful to devise a single parametric function that provides an overall measure of these differences and that can be readily estimated from the data. One such function is $[(1/t)\sum_{i=1}^{t} \tau_i^2]^{1/2}$, where $\tau_i = \alpha_i - \beta_i - (1/t)\sum_{j=1}^{t}(\alpha_j - \beta_j)$ (which is an estimable linear combination of the parameters of Model 3 that is interpretable as the difference between the home-court advantage for the ith team and the overall home-court advantage). This function is interpretable as the root mean squared deviation of the team-specific home-court advantages $\alpha_1 - \beta_1, \ldots, \alpha_t - \beta_t$ from the overall home-court advantage. It can be estimated by the statistic $\left[\max\{0, (1/t)\sum_{i=1}^{t}[\hat{\tau}_i^2 - \hat{v}_i]\}\right]^{1/2}$, where $\hat{\tau}_i$ is the least squares estimator of τ_i, and $\hat{v}_i$ is the usual unbiased estimator of $\text{var}(\hat{\tau}_i)$. The motivation for regarding this statistic as an estimator of $[(1/t)\sum_{i=1}^{t} \tau_i^2]^{1/2}$ comes from observing that $(1/t)\sum_{i=1}^{t}[\hat{\tau}_i^2 - \hat{v}_i]$ is an unbiased estimator of $(1/t)\sum_{i=1}^{t} \tau_i^2$.

For our data, the estimated value $[(1/t)\sum_{i=1}^{t} \tau_i^2]^{1/2}$ is 0.59. By comparison, the least squares estimate of the overall home-court advantage $(1/t)\sum_{i=1}^{t}(\alpha_i - \beta_i)$ equals 4.70, indicating that team-to-team variability in the home-court advantage may not be of great practical importance.

Similar conclusions about the importance of the overall home-court advantage and team-to-team differences in the home-court advantage can be reached by comparing the coefficients of determination for Models 1–3, which are, respectively, $R_1^2 = S_1 \div S_t$, $R_2^2 = (S_1 + S_{2|1}) \div S_t$, and $R_3^2 = (S_1 + S_{3|1}) \div S_t$. (It seems more appropriate to define the coefficients of determination for Models 1–3 in terms of uncorrected [for the mean] sums of squares than in terms of corrected sums of squares.)

For our data, $R_1^2 = .475$, $R_2^2 = .557$, and $R_3^2 = .606$. In comparing these three values, note that the increase in the coefficient of determination from .475 to .557 is achieved with the addition of only one parameter to the model, whereas the further increase from .557 to .606 requires the addition of another 154 parameters.

Table 2. Overall Relative Performance Levels and Home-Court Advantages for Selected Teams

Rank on overall relative performance level: Model 3	*Team*	*Overall relative performance level: Model 3*	*Overall relative performance level: Model 2*	*Home-court advantage*	*Rank on home-court advantage*
1	Duke	21.1	21.2	12.0	12
2	Indiana	20.4	20.8	13.0	8
3	Kansas	19.1	19.5	6.6	50
4	Arizona	17.2	17.4	11.2	15
5	Ohio State	14.9	15.2	2.4	108
6	Cincinnati	13.8	14.6	−1.6	139
7	Oklahoma State	13.4	13.9	11.2	16
8	Kentucky	12.5	12.3	5.9	63
9	Missouri	12.3	13.8	−0.9	136
10	Oklahoma	12.2	13.3	1.2	124
11	Arkansas	11.9	12.4	2.1	114
12	Massachusetts	11.1	11.2	6.2	57
13	North Carolina	10.7	11.0	5.5	72
14	St. John's	10.3	9.8	9.3	27
15	Iowa	10.3	10.0	4.3	88
16	Nevada–Las Vegas	10.0	11.1	6.0	62
17	Louisiana State	9.7	9.9	2.2	110
18	Michigan	9.6	9.4	−0.4	134
19	UCLA	9.6	10.3	−5.4	149
20	Michigan State	9.4	9.3	3.7	96
35	Rhode Island	6.4	6.1	12.5	10
48	Villanova	4.6	3.3	14.7	2
67	Butler	2.4	2.7	−9.7	155
77	Utah	0.7	1.0	−6.2	151
83	Rice	−0.2	1.3	−7.8	153
88	Wisconsin	−0.7	−1.2	14.1	4
92	Clemson	−1.5	−1.6	13.9	6
102	South Alabama	−2.7	−2.6	−8.0	154
105	Pacific	−3.4	−2.4	−4.4	146
106	New Orleans	−3.6	−2.7	−6.3	152
111	Kansas State	−4.5	−3.0	15.8	1
119	Loyola (IL)	−5.3	−6.0	12.7	9
120	St. Joseph's (PA)	−5.7	−5.4	−5.5	150
124	Miami (FL)	−6.1	−5.8	−5.0	148
126	Indiana State	−6.4	−6.6	13.8	7
128	Cal. St. Northridge	−6.7	−8.9	13.9	5
134	Eastern Michigan	−9.6	−10.0	−4.7	147
154	Chicago State	−22.1	−23.1	14.7	3

It is of some interest to identify those teams that had exceptionally high or exceptionally low home-court advantages and to ascertain whether there is any tendency for the better teams to have home-court advantages that are above (or below) average. Table 2 and Figure 1 are helpful in that regard.

Table 2 lists for selected values of i the ith team's overall relative performance level as determined from Model 2 (the least squares estimate of $\beta_i - (1/t)\sum_{j=1}^{t}\beta_j$), the ith team's overall relative performance level as determined from Model 3 (the least squares estimate of $\frac{1}{2}[\alpha_i + \beta_i - (1/t)\sum_{j=1}^{t}(\alpha_j + \beta_j)]$), and the ith team's home-court advantage as determined from Model 3 (the least squares estimate of $\alpha_i - \beta_i$). The teams included in Table 2 are those that were among the top 20 in overall relative performance level as determined from either Model 2 or Model 3 and those that were among either the top 10 or bottom 10 in home-court advantage. The ordering of the teams in Table 2 is by overall relative performance level (as determined from Model 3).

Figure 1 displays (in the form of a scatterplot) the home-court advantages and the overall relative performance levels (as determined from Model 3) for the 155 teams. Judging from the evidence provided by Table 2 and Figure 1, there does not appear to be any substantial positive or negative relationship between home-court advantage and overall performance level. For example, Duke and Indiana, which were the two strongest teams, had the 12th and 8th largest home-court advantages; however, Chicago State, which was among the very weakest teams, had an even larger home-court advantage.

Any of the three models could conceivably be used as a basis for predicting the difference in score, say y_*, for a future game. The least squares estimator, say $\hat{y}_*$, of the expected difference in score $\mathrm{E}(y_*)$ can be regarded as a point predictor of y_*; in fact, it is the best (minimum mean squared error) linear unbiased predictor. Moreover, a $100(1-\alpha)\%$ prediction interval is provided by the interval with end points $y_* \pm t_{\alpha/2}(\hat{\sigma}^2 + \hat{v}_*)^{1/2}$, where $\hat{\sigma}^2$ and $\hat{v}_*$ are the usual unbiased estimators of σ^2 and $\mathrm{var}(\hat{y}_*)$,

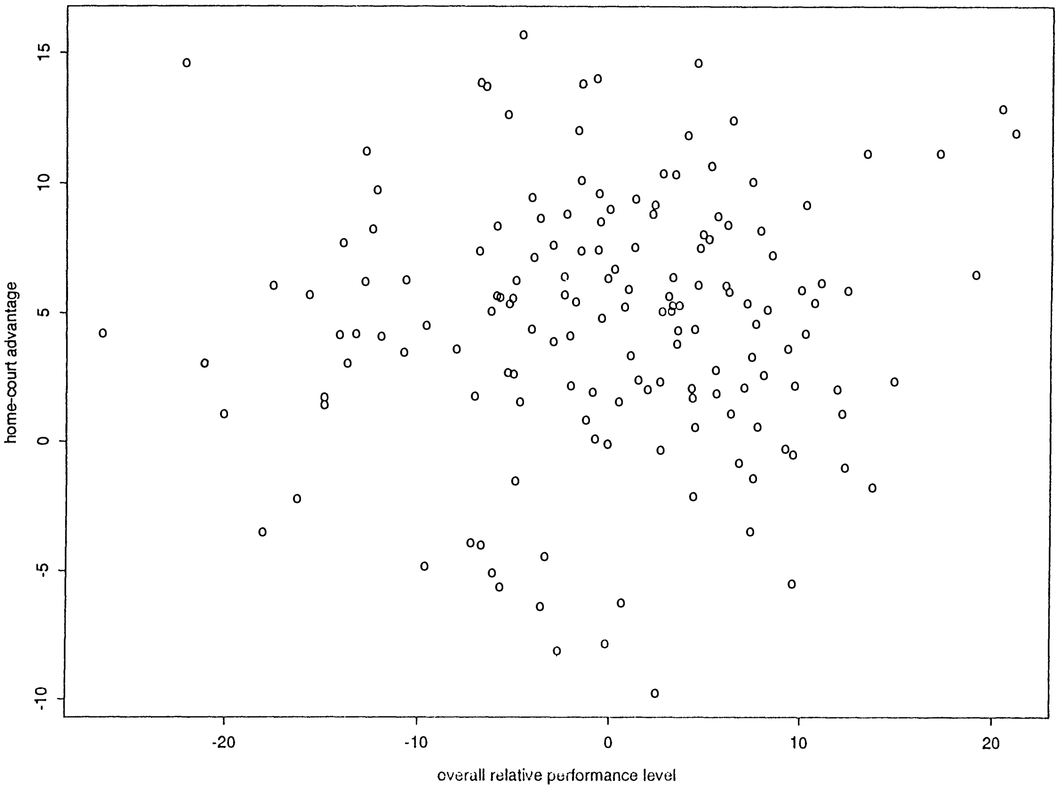

Figure 1. Scatterplot of home-court advantage versus overall relative performance level.

respectively, and where $t_{\alpha/2}$ is the upper-$\alpha/2$ point of Student's distribution with (depending on the choice of model) $n-t+1$, $n-t$, or $n-2t+1$ df. Note that this interval differs from the interval with end points $\hat{y}_* \pm t_{\alpha/2}\hat{v}_*^{1/2}$, which is a $100(1-\alpha)\%$ confidence interval for $E(y_*)$. Our results on the home-court advantage strongly suggest that the predictions derived from Model 2 will tend to be more accurate than those derived from Model 1 or Model 3.

The point and interval predictions for y_* can be highly sensitive to the choice of model and (in the case of Models 2 and 3) to the site of the game. To illustrate this sensitivity, we used our data to predict the difference in score for a game between Oklahoma State (Team 8) and Michigan (Team 31). We chose a game between these two teams to obtain an extreme case in which the home-court advantage of one of the teams (Oklahoma State) was estimated to be relatively large and that of the other was estimated to be relatively small. An additional motivation for choosing this game was that Oklahoma State and Michigan played each other (on a neutral court) in the third round of the postseason Division I men's basketball tournament (with Oklahoma State losing to Michigan by 4 points). The various predictions are listed in Table 3. Note in particular that the prediction (and confidence) intervals derived from Model 2 are somewhat shorter than those derived from Models 1 and 3.

In Section 2, for any game played on a neutral court we adopted the notational convention of arbitrarily labeling one of the two opposing teams the home team. Note that none of our results is affected by the arbitrariness of the labeling. To see this, suppose that the ith and jth teams played a game on a neutral court and that the ith team was labeled the home team. If the labels were switched (so that the jth team were labeled the home team), then the only effect on the vectors $\mathbf{y}$, $\mathbf{x}$, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_t$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t$, and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t$ would be to change the sign (and location) of the ijkth element of each vector. Such a change would not alter the total sum of squares $\mathbf{y}'\mathbf{y}$ or the normal equations for Models 1–3, and hence (in light of the nature of the statistical procedures we have used) would not affect any of our results.

5. DISCUSSION

By adopting Models 2 and 3, we are able to use relatively standard (fixed-effects) linear-model methods to make inferences about the home-court advantage and about team-to-team differences in the home-court advantage (as discussed in Sections 3 and 4). We can attempt to improve on Models 2 and 3 by introducing various modifications and refinements, though such changes may necessitate the use of more sophisticated statistical methods. Some possible changes are as follows.

Change 1. Instead of supposing that every team has the same home-court advantage (as in Model 2) or allowing for the possibility that every team has a different home-court advantage (as in Model 3), we could adopt an "intermediate" approach in which the teams are divided into two or more classes, and the home-court advantage is assumed to be the same for any two games in which both home teams are from the same class and both visiting teams are from the same class but is allowed to differ for games in which either the home teams are from different classes or the visiting teams are from different classes.

Table 3. Predictions or Estimates for the Difference in Score or Expected Difference in Score Between Oklahoma State and Michigan in a Game Played on Either Home Court or on a Neutral Court

Court and assumed model	Expected difference in score	Point estimate or prediction	95% confidence interval for expected difference	95% prediction interval
Oklahoma State				
1	$\beta_8 - \beta_{31}$	4.8	[−2.0, 11.5]	[−19, 29]
2	$\lambda + \beta_8 - \beta_{31}$	9.2	[3.0, 15.4]	[−13, 31]
3	$\alpha_8 - \beta_{31}$	9.2	[0.5, 17.9]	[−13, 32]
Michigan				
1	$\beta_8 - \beta_{31}$	4.8	[−2.0, 11.5]	[−19, 29]
2	$-\lambda + \beta_8 - \beta_{31}$	−0.2	[−6.4, 6.1]	[−22, 22]
3	$\beta_8 - \alpha_{31}$	−1.6	[−10.1, 6.8]	[−24, 21]
Neutral				
1	$\beta_8 - \beta_{31}$	4.8	[−2.0, 11.5]	[−19, 29]
2	$\beta_8 - \beta_{31}$	4.5	[−1.7, 10.7]	[−17, 26]
3	$\beta_8 - \beta_{31}$	−2.0	[−10.8, 6.7]	[−25, 21]

The classes could be formed on the basis of home attendance, altitude, and other relevant variables (see Sallas and Harville 1988).

Change 2. The home-court advantage could be allowed to vary from game to game as a (linear or nonlinear) function of the total number of points scored by the two opposing teams.

Change 3. The home-court advantage λ in Model 2 could be replaced by $(\lambda_i + \lambda_j)/2$, where $\lambda_1, \ldots, \lambda_t$ are random effects with common unknown mean λ and common unknown variance σ^2_λ

Change 4. The quantities $\beta_1, \ldots, \beta_t$ (in Model 2 or 3) and the quantities $\alpha_1, \ldots, \alpha_t$ (in Model 3) could be treated as random effects rather than as unknown parameters and could possibly be allowed to vary with time (see Harville 1977).

Change 5. A (fixed or random) interaction effect, say γ_{ij} (possibly restricted so that $\gamma_{ji} = -\gamma_{ij}$), could be added to the model equation for y_{ijk}.

Change 6. The assumptions about the e_{ijk}'s (the residual effects) could be relaxed to allow, for instance, for the possibility that the outcomes of games involving the same team or teams are correlated to an extent that diminishes with the intervening time (although the results of Harville [1977], which are for football data, suggest that any such correlations may be small and relatively unimportant).

Ultimately, the importance of accounting for the home-court advantage in the rating of teams (or of accounting for the presence of team-to-team differences in the home-court advantage or for the presence of other factors) depends on the underlying objectives. If the ratings were to be used in the selection or seeding of teams for the NCAA Division I postseason tournament, then they should accurately reflect the expected relative performance levels of the various teams in games played on a neutral court—though as discussed by Harville (1977, 1978), accuracy might not be the only consideration.

One way to compare alternative rating systems with regard to accuracy is to adopt (for that purpose) a suitably flexible model, to formally define accuracy in terms of that model, and to then evaluate the alternative systems (resorting if necessary to simulation) on the basis of that criterion. Another approach would be to apply the alternative rating systems retrospectively to the regular (pre-tournament) part of each of several previous seasons, to translate the alternative ratings into predictions about the tournament games, and to compare the alternative systems empirically on the basis of those predictions.

[*Received July 1992. Revised February 1993.*]

REFERENCES

Harville, D. A. (1977), "The Use of Linear-Model Methodology to Rate High School or College Football Teams," *Journal of the American Statistical Association*, 72, 278–289.

——— (1978), "Football Ratings and Predictions via Linear Models," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 74–82.

National Collegiate Athletic Association (1991), *Official 1992 NCAA Basketball*, Overland Park, KS: Author.

Sallas, W. M., and Harville, D. A. (1988), "Noninformative Priors and Restricted Maximum Likelihood Estimation in the Kalman Filter," in *Bayesian Analysis of Times Series and Dynamic Models*, ed. J. C. Spall, New York: Marcel Dekker, pp. 477–508.

Searle, S. R. (1971), *Linear Models*, New York: John Wiley.

Stefani, R. T. (1977), "Football and Basketball Predictions Using Least Squares," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-7, 117–121.

——— (1980), "Improved Least Squares Football, Basketball, and Soccer Predictions," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-10, 116–123.