# Skill, Luck, and Streaky Play on the PGA Tour

Robert A. CONNOLLY and Richard J. RENDLEMAN JR.

In this study, we implement a random-effects model that estimates cubic spline-based time-dependent skill functions for 253 active PGA Tour golfers over the 1998–2001 period. Our model controls for the first-order autocorrelation of residual 18-hole scores and adjusts for round–course and player–course random effects. Using the model, we are able to estimate time-varying measures of skill and luck for each player in the sample. Estimates of spline-based player-specific skill functions provide strong evidence that the skill levels of PGA Tour players change through time. Using the model, we are able to rank golfers at the end of 2001 on the basis of their estimated mean skill levels and identify the players who experienced the most improvement and deterioration in their skill levels over the 1998–2001 period. We find that some luck is required to win PGA Tour events, even for the most highly skilled players. Player–course interactions contribute very little to variability in golfer performance, but the particular course rotations to which players are assigned can have a significant effect on the outcomes of the few tournaments played on more than one course. We also find evidence that a small number of PGA Tour participants experience statistically significant streaky play.

KEY WORDS: Golf; Hot hands; Luck; Random effects; Skill; Smoothing spline.

## 1. INTRODUCTION

Like all sports, outcomes in golf involve elements of both skill and luck. Perhaps the highest level of skill in golf is displayed on the PGA Tour. Even among these highly skilled players, however, a small portion of each 18-hole score can be attributed to luck.

In this article, we model the development of skill over time for PGA Tour players, and we offer a decomposition of the nonskill portion of their golf scores. We estimate the properties of each source of variability and highlight their temporal dependence, thus connecting our work to the hot-hands literature.

Our work covers the 1998–2001 period. We decompose individual golfer's scores into skill-based and unexpected components by using Wang's (1998) smoothing spline model to estimate each player's mean skill level as a function of time, while simultaneously estimating the correlation in the random error structure of fitted player scores and the random effects associated with round–course and player–course interactions. The fitted spline values of this model provide estimates of expected player scores after controlling for these error adjustments. We use the fitted values from this model to characterize the evolution of skill changes among PGA Tour golfers and use model residuals as the basis for our study of luck and streaky play. This decomposition has an antecedent in Klaassen and Magnus (2001) who modeled the probability of winning a tennis point as a function of player quality (proxied by player ranking), context (situational) variables, and a random component. They noted that failing to account for quality differences will create pseudo-dependence in scoring outcomes, because winning reflects, in part, player quality, and player quality is generally persistent. Parameter estimates from their dynamic, panel random-effects model using four years of data from Wimbledon matches suggest that the iid hypothesis for tennis points is a good approximation, provided there are controls for player quality.

Our tests show that temporal changes in skill are common among PGA Tour players, and for some players, the relationship between skill and time is highly nonlinear. We find that the random effects associated with round–course interactions, which capture the relative difficulty of each round, can be substantial, ranging from approximately $-4$ to $+7$ strokes per round. Although we find that it is also important to adjust player scores for the random effects of player–course interactions, these effects tend to be dramatically smaller. We show PGA Tour players cannot win tournaments without experiencing some good luck, even after accounting for skill changes. On average, it took 9.6 strokes of cumulative "good luck" to win a tournament during our sample period. We find evidence of significant positive first-order autocorrelation in the error structure for approximately 9% of the golfers in our sample. We show that most of these significant autocorrelations do not represent false discoveries and, therefore, conclude that a limited set of PGA Tour players actually experience episodes of statistically significant streaky play. However, after removing the effects of first-order autocorrelation from the residual scores, we find little additional evidence of streakiness. This suggests that any statistical study of sporting performance should estimate skill dynamics and deviations from normal performance while simultaneously accounting for the relative difficulty of the task and the potential autocorrelation in unexpected performance outcomes.

The remainder of the article is organized as follows. In Section 2, we describe our data and criteria for including players in our statistical samples. In Section 3, we present our model of skill and luck and related specification tests. Using this model, we describe the evolution of player skill in Section 4. In Section 5, we decompose the variability in player scores into their components, discuss the relationship between variability and luck, and demonstrate how much luck it takes to win a PGA Tour event. In Section 6, we address the issue of streaky play as it relates to golf. A final section provides a summary and concluding comments.

## 2. DATA

We have collected individual 18-hole scores for every player in every stroke play event on the PGA Tour for years 1998–

Robert A. Connolly is Associate Professor (E-mail: *Robert_Connolly@unc.edu*) and Richard J. Rendleman Jr. is Professor (E-mail: *richard_rendleman@unc.edu*), Kenan-Flagler Business School, University of North Carolina, Chapel Hill, NC 27599. We thank Tom Gresik and seminar participants at the University of Notre Dame for helpful comments, along with Carl Ackermann and David Ravenscraft who provided significant input and assistance during the early stages of this study. We also thank Mustafa Gültekin for computational assistance and Yuedong Wang and Douglas Bates for assistance in formulating, programming and testing portions of our estimation procedures. We are particularly grateful to the editor, associate editor, and referees who all provided very useful comments.


Journal of the American Statistical Association
March 2008, Vol. 103, No. 481, Applications and Case Studies
DOI 10.1198/016214507000000310

2001 for a total of 76,456 scores distributed among 1,405 players. (All but two regular PGA Tour events involve stroke play. A stroke play event is one in which all player strokes must be recorded and count equally over the entire tournament. The player with the lowest total score wins.) Our data include all stroke play events for which participants receive credit for earning official PGA Tour money, even though some of the events, including all four "majors," are not actually run by the PGA Tour. The data were collected, primarily, from

- *www.pgatour.com*,
- *www.golfweek.com*,
- *www.golfonline.com*,
- *www.golfnews.augustachronicle.com*,
- *www.insidetheropes.com*, and
- *www.golftoday.com*.

When we were unable to obtain all necessary data from these sources, we checked national and local newspapers, and, in some instances, contacted tournament headquarters directly.

Our data cover scores of players who made and missed cuts. [Although there are a few exceptions, after the second round of a typical PGA Tour event, the field is reduced (cut) to the 70 players, including ties, with the lowest total scores after the first two rounds.] The data also include scores for players who withdrew from tournaments and who were disqualified; as long as we have a score, we use it. We also gathered data on where each round was played. This is especially important for tournaments such as the Bob Hope Chrysler Classic and ATT Pebble Beach National Pro-Am played on more than one course.

The great majority of the players represented in the sample are not regular members of the PGA Tour. Nine players in the sample recorded only one 18-hole score over the 1998–2001 period, and 565 players recorded only two scores. Most of these 565 players qualified to play in a single U.S. Open, British Open, or PGA Championship, missed the cut after the first two rounds, and subsequently played in no other PGA Tour events.

Note that 1,069 players, 76.1% of the players in the sample, recorded 50 or fewer 18-hole scores. This same group recorded 5,895 scores, which amounts to only 7.7% of the total number of scores in the sample. In addition, 1,162 players, representing 82.7% of the players in the sample, recorded 100 or fewer scores. This group recorded 12,849, or 16.8% of the sample. The greatest number of scores recorded by a single player was 457 recorded by Fred Funk.

In estimating the skill levels of individual players, we employ the smoothing spline model of Wang (1998), which adjusts for correlation in random errors. Simulations by Wang indicate that 50 sample observations is probably too small and that approximately 100 or more observations are required to obtain dependable statistical estimates of cubic spline–based mean estimates. After examining player names and the number of rounds recorded by each player, we have concluded that a sample of players who recorded more than 90 scores is reasonably homogeneous and likely to meet the minimum sample size requirements of the cubic spline methodology. If we were to include players who recorded fewer scores, we would add a mix of established senior players (e.g., Hale Irwin), relatively inactive middle-aged players (Jerry Pate), "old-timers" (Arnold Palmer), up-and-coming stars (Chad Campbell), established European Tour players (Robert Karlsson), and many generally unknown players. Because this group is not representative of typical PGA Tour participants, excluding them minimizes potential distortions in estimating the statistical properties of golf scores of regular players on the Tour. After limiting our sample to players who recorded more than 90 18-hole scores, the resulting sample consists of 64,364 observations.

## 3. MODEL OF SKILL AND LUCK

### 3.1 Overview

We organize our model of golfer skill and luck using the following general structure:

$$\mathbf{s} = \mathbf{P} f(\bullet) + \mathbf{R} b_2 + \mathbf{C} b_3. \tag{1}$$

In (1), $\mathbf{s} = (s_1, \ldots, s_{253})'$ is an $N = 64{,}364$ vector of 18-hole scores, subdivided into 253 player groups of $n_i$ scores per group, with $N = \sum_{i=1}^{253} n_i$. Within each group, the scores are ordered sequentially, with $s_i = (s_{i1}, \ldots, s_{in_i})'$ denoting the vector of scores for player $i$ ordered in the chronological sequence $g_i = 1, 2, \ldots, n_i$. We refer to $g_i$ as the sequence of player $i$'s "golf times." The usual error term is subsumed in $f(\bullet)$.

$\mathbf{P}$ is an $N \times 253$ matrix that identifies the player, among the 253, associated with each score. $f(\bullet) = (f_1(\bullet), \ldots, f_{253}(\bullet))'$ is a vector of 253 player-specific cubic spline functions estimated via Wang's (1998) model (described in more detail in the next section). Therefore, $\mathbf{P} f(\bullet)$ captures time variation in skill for each of the 253 golfers.

We assume there are two important sources of golf-related random effects, one related to round–course interactions and the other related to player–course interactions. We identify round–course interactions associated with each score using $\mathbf{R}$, an $N \times 848$ matrix. A round–course interaction is defined as the interaction between a regular 18-hole round of play in a specific tournament and the course on which the round is played. For 157 of 182 tournaments in our sample, only one course is used, and, therefore, there is only one such interaction per round. By contrast, the first four rounds of the Bob Hope Chrysler Classic are played on four different courses using a rotation that assigns each tournament participant to each of the four courses over the first four days. A cut is made after the fourth round, and a final round is played the fifth day on a single course. Thus, the Bob Hope tournament consists of 17 round–course interactions—four for each of the first four days of play and one additional interaction for the fifth and final day. There are a total of 848 round–course interactions in our sample and, on average, 254 such interactions per player. $b_2$ is an 848-dimensional vector of estimated random effects associated with each of the 848 round–course interactions.

We identify player–course interactions associated with each score using $\mathbf{C}$, an $N \times 12{,}485$ matrix of 253 groups of nested player–course interactions. $b_3 = (b_{3\,1}, \ldots, b_{3\,253})'$ is the vector of nested random player–course effects grouped by player, with $b_{3j} = (b_{3j1}, \ldots, b_{3jm_j})'$, where $m_j$ is the total number of nested player–course interactions associated with player $j$.

### 3.2 Random Effects With Constant Player Skill

Because estimating the spline portion of our model is computationally intensive (the model takes approximately 40 hours to estimate on a Windows XP–based PC with 1 GB RAM using a 2.80-GHz Intel® Xeon™ processor and over five days to compute bootstrap test statistics), we employ simpler versions of the skill portion of the model to identify the appropriate golf-related random effects. Specifically, we substitute $b_1 + \boldsymbol{\varepsilon}$ for the $f(\bullet)$ vector of player-specific cubic spline functions, where $b_1 = (b_{1\,1}, \ldots, b_{1\,253})'$ , a 253-dimensional vector of random player effects, and $\boldsymbol{\varepsilon}$ is a vector of $N$ normally distributed random errors with zero mean and constant variance.

From a sequence of $F$ and log-likelihood tests, we draw the following conclusions about the appropriate random-effects specification: (a) $\mathbf{s} = \mathbf{P}b_1 + \mathbf{R}b_2 + \mathbf{C}b_3 + \boldsymbol{\varepsilon}$ is a superior specification to (b) $\mathbf{s} = \mathbf{P}b_1 + \mathbf{R}b_2 + \boldsymbol{\varepsilon}$ [$\chi^2(1) = 25.084$, $p$ value $=$ .000], which indicates that it is important to include player–course effects along with round–course effects. Separately, in specification (a), using the Akaike information criterion (AIC), we also determine that the best way to control for the random effects associated with the relative difficulty of each round is through random round–course interaction effects rather than round, tournament, or course effects.

### 3.3 Random Effects With Time-Varying Skill

In the aforementioned tests, we assume that each player's expected score is constant over the entire four-year sample period. To account for potential time variation, without actually estimating the full spline model, we employ the following method for specifying time variation in individual player skill. If a full calendar year has passed, at least 25 scores were used in the estimation of the random effects associated with the player's basic skill level, and at least 25 additional scores were recorded for the same player, then a new incremental (random) skill effect for that player is estimated. For players who participated actively in all four years of the sample, this procedure results in the estimation of random skill effects in each of the four calendar years. In this specification, $\mathbf{P}_2$ is an $N \times 245$ matrix that connects players and scores for the 245 players whose scores cover two or more years, $\mathbf{P}_3$ is an $N \times 196$ matrix that connects players and scores for the 196 players whose scores cover three or more years, and $\mathbf{P}_4$ is an $N \times 128$ matrix that connects players and scores for the 128 players whose scores cover four years. $b_{p2}, b_{p3}$, and $b_{p4}$ are vectors of estimated random effects associated with these same player groups. Although the 25-score criterion is somewhat arbitrary, it is not critical, because the model we actually use for estimating player scores is based on Wang's smoothing spline methodology and does not use the 25-score criterion.

In another series of tests, we show that (c) $\mathbf{s} = \mathbf{P}b_1 + \mathbf{P}_2b_{p2} + \mathbf{P}_3b_{p3} + \mathbf{P}_4b_{p4} + \mathbf{R}b_2 + \mathbf{C}b_3 + \boldsymbol{\varepsilon}$ is a superior model specification relative to (a) $\mathbf{s} = \mathbf{P}b_1 + \mathbf{R}b_2 + \mathbf{C}b_3 + \boldsymbol{\varepsilon}$ [$\chi^2(3) = 341.79$, $p$ value $=$ .000]. In addition, we examine the same combination of golf-related random effects referred to in Section 3.2. Using the AIC, we conclude that the best way to model the random effects associated with relative round difficulty is through round–course effects and that player–course effects should be included with round–course effects. It should be noted that Berry's (2001) model for predicting player scores is equivalent to (a) using random effects for actual rounds rather than for round–course interactions.

We recognize that allowing player skill to change only by (approximate) calendar year is a very restrictive approach to modeling time variation in skill. Therefore, in the final version of our model, we allow the time pattern of skill changes to be determined statistically using Wang's smoothing spline. We now describe how we employ Wang's model in our estimation of time-dependent player skill.

### 3.4 Skill Model (Wang's Smoothing Spline)

Wang's smoothing spline model, as applied to individual player $i$, is expressed as follows:

$$f_i(\bullet) = h_i(g_i) + \boldsymbol{\theta}_i = h_i(g_i) + \boldsymbol{\varphi}_i + \boldsymbol{\eta}_i. \tag{2}$$

In (2), $h_i(g_i)$ is Wang's smoothing spline function applied to player $i$'s golf scores over his specific golf times $g_i = 1, 2, \ldots, n_i$. (As noted earlier, $g_i$ represents a counting of player $i$'s golf scores in chronological order.) $\boldsymbol{\theta}_i$ is a vector of potentially autocorrelated random errors associated with player $i$'s spline fit, with $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{in_i})' \sim \mathrm{N}(0, \sigma_i^2\mathbf{W}_i^{-1})$ and $\sigma_i^2$ unknown. In Wang's model, $\mathbf{W}_i^{-1}$ is a covariance matrix whose form depends on specific assumptions about dependencies in the errors, for example, first-order autocorrelation for time series, compound symmetry for repeated measures, and so on. See Wang (1998, p. 343) for further details.

For any given player $i$, let $\mathbf{f} = (f_1, \ldots, f_n)'$ denote the vector of the player's $n$ sequentially ordered golf scores, net of estimated round–course and player–course effects. Also, let $\mathbf{h} = (h(t_1), \ldots, h(t_n))'$ denote the vector of the player's estimated cubic spline function evaluated at points $t_1, \ldots, t_n$, which represent golf times $g = 1, 2, \ldots, n$ scaled to the [0, 1] interval. In Wang's model, as applied here, for each player, one chooses the cubic spline function $h(t)$, the smoothing parameter $\lambda$, and the first-order autocorrelation coefficient $\phi$ embedded in $\mathbf{W}$ that minimizes $\frac{1}{n}(\mathbf{f}-\mathbf{h})'\mathbf{W}(\mathbf{f}-\mathbf{h}) + \lambda \int_0^1 (d^2h(t)/dt^2)^2\,dt$. "The parameter $\lambda$ controls the trade-off between goodness-of-fit and the smoothness of the [spline] estimate" (Wang 1998, p. 342).

In (2), we break $\boldsymbol{\theta}_i$ into two parts, $\boldsymbol{\varphi}_i + \boldsymbol{\eta}_i$, where $\boldsymbol{\varphi}_i$ represents the autocorrelated component of $\boldsymbol{\theta}_i$ and $\boldsymbol{\eta}_i$ is assumed to be white noise. Inasmuch as there are likely to be gaps in calendar time between some adjacent points in a player's golf time, it is unlikely that random errors around individual player spline fits follow higher order autoregressive processes [i.e., AR($n$), $n > 1$]. Therefore, we assume that each $\boldsymbol{\theta}_i$ follows a player-specific AR(1) process with first-order autocorrelation coefficient $\phi_i$.

Our estimation model represents a generalized additive model with random effects $\boldsymbol{\eta}_i$, $b_2$, and $b_3$. The methodology for estimating this model is described in the Appendix. (In fitting the model, we compute a pseudo adjusted $R^2$ of .296, calculated as 1 − Mean square error/Mean square total.)

### 3.5 Specification Tests

To check for possible model misspecification in connection with the AR(1) assumption, we compute the Ljung–Box (1978) $Q$ statistic associated with the $\eta$ errors for each player for lags equal to the minimum of 10 or 5% of the number of rounds played. Ljung (1986) suggested that no more than 10 lags should be used for the test, and Burns (2002) recommended that the number of lags should not exceed 5% of the length of the series. Only 6 of 253 $Q$ statistics are significant at the 5% level. When the same test is applied to $\theta$ errors, which may be serially correlated, 17 $Q$ statistics are significant.

To guard against inappropriate conclusions in summarizing test statistics across multiple golfers, we apply the false discovery rate (FDR) method proposed by Storey (2002, 2003) to both sets of tests. The minimum chance of a false positive discovery, that is, Storey's $q$ value, among all 253 FDRs associated with $\eta$ errors is .999. This suggests that our first-order autocorrelation correction leaves white-noise errors for all 253 players. The following table summarizes the $p$ values and $q$ values for the 17 players for whom the Ljung–Box $Q$ statistic as applied to $\theta$ errors is significant.

| Players | $p$ value | $q$ value |
|---|---|---|
| 1–8 | .006–.012 | .391 |
| 9–11 | .016–.019 | .440 |
| 12 | .026 | .557 |
| 13 | .032 | .627 |
| 14–15 | .037–.038 | .645 |
| 16–17 | .045–.048 | .712 |

Although all the $q$ values are relatively high, some of the $p$ values are almost certainly significant, not just false discoveries. Therefore, on the basis of these tests, we conclude that it is important to take potential first-order autocorrelation into account when estimating individual players' cubic spline functions and that no higher order autocorrelation adjustments are necessary. In the next section we provide further analysis of potential autocorrelation in the $\theta$ errors.

We employ the bootstrap to test individual player spline fits against alternative specifications of player skill. All bootstrap tests are based on balanced sampling of 40 samples per player. Although 40 samples is well below the number required to estimate precise confidence intervals for each individual player, a total of $40 \times 253 = 10{,}120$ bootstrap samples are taken over all 253 players, requiring over five days of computation time. The 10,120 total bootstrap samples should be more than sufficient to draw general inferences about statistical significance in this setting.

Wang and Wahba (1995) described how the bootstrap can be used in connection with smoothing splines that are estimated without taking the autocorrelation in residual errors into account. We modify their method so that the bootstrap samples are based on $\eta$ residuals, which adjust predicted scores for autocorrelation in prior $\theta$ residuals.

Figure 1 summarizes $p$ values associated with bootstrap tests of the 253 cubic spline fits against the following alternative methods of estimating a player's 18-hole score (after subtracting random round–course effects):

- The player's mean score
- The player's mean score in each approximate calendar year period as defined in Section 3.3
- The player's score estimated as a linear function of time
- The player's score estimated as a quadratic function of time.

For each of the four tests, we form a test statistic

$$\hat{\zeta} = \frac{\text{RSE(alt)}}{n - df_{\text{alt}}} - \frac{\text{RSE(spline)}}{n - df_{\text{spline}} - 1},$$

where $n$ is the number of 18-hole scores for a given player, $df_{\text{alt}}$ and $df_{\text{spline}}$ are the number of degrees of freedom associated with the estimation of the alternative and spline models, respectively, and RSE(alt) and RSE(spline) are total residual squared errors from the alternative and spline models. We subtract 1 in the denominator of the second term to account for the additional degree of freedom associated with the estimation of the first-order autocorrelation coefficient. For the purposes of this test, RSE(spline) is based on $\eta$ errors, because these errors reflect the complete information set derived from each spline fit. The test statistic $\hat{\zeta}$ is suggested by Efron and Tibshirani [1998, pp. 190–192, 200 (problem 14.12)] for testing the predictive power of an estimation model formulated with two different sets of independent variables. The test statistic is computed for each of 40 bootstrap samples per player.

We employ the following interpolation method for estimating $p$ values for each of the $\hat{\zeta}$ test statistics. For each player, we sort the test statistics in ascending order. The first becomes the cutoff point for the probability interval from 0 to $1/40 = .025$, the second becomes the cutoff point for the interval 0 to $2/40 = .05$, and so on. We use linear interpolation to estimate $p$ values that fall between these intervals. For example, if the first of the ordered test statistic is $-.1$ and the second is .3, we compute the probability of a coefficient less than 0 as .03125. (We also experimented with various cubic spline and piecewise polynomial spline-based interpolations, but these gave irrational results in some cases.) If all bootstrap estimates of the test statistic are positive, we set the $p$ value to $\frac{1}{2}$ of .025, or .0125. If all are negative, we set the $p$ value to 0.98875. The distribution of these $p$ values for all four tests are shown in Figure 1.

The spline model is significantly superior (at the 5% level) to the player's mean score for 71 of the 253 players in the sample (approximately 28% of the players). It is significantly superior to a linear time trend for 25 of the players (10%), to the mean score computed for each (approximate) calendar year for 13 players (5%), and to a quadratic time trend for 10 players (4%). It should be noted, however, that if a constant mean, linear, or quadratic function is the appropriate form to capture skill variation for a particular player, the spline model should be sufficiently flexible to reproduce this same functional form. Moreover, if a more complex relationship between time and skill, such as the approximate calendar year model, is appropriate, the spline model should provide a smoothed version of such a relationship as well. Therefore, even if a simple linear relationship between time and skill were the appropriate functional form for 252 of 253 players, we would be better served by the spline model, because it should reflect the same linear form for the 252 and provide an approximation to the appropriate functional form for the 253rd player.

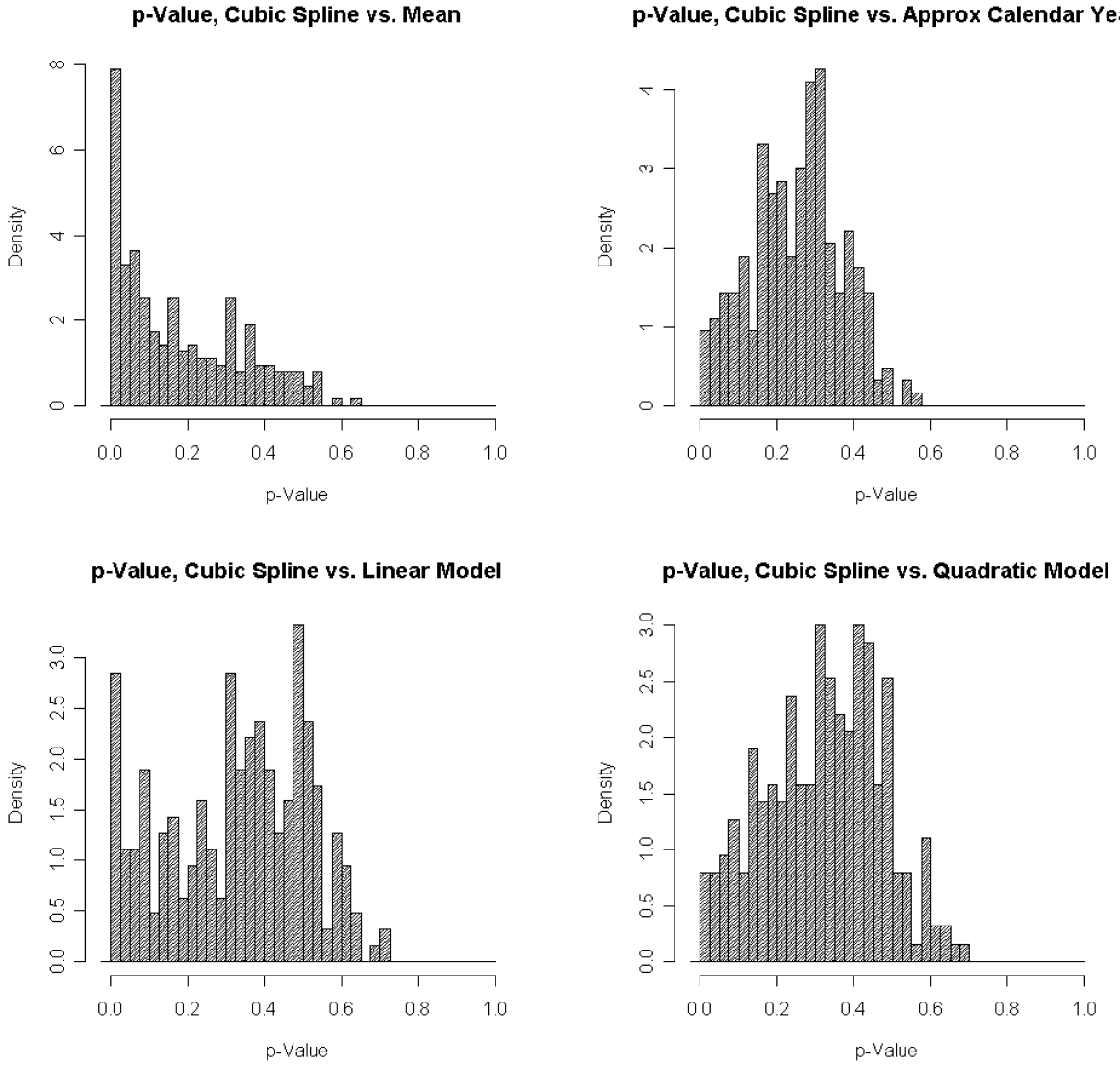


Figure 1. $p$ values associated with bootstrap tests of cubic spline fits versus alternative models of time variation in skill.

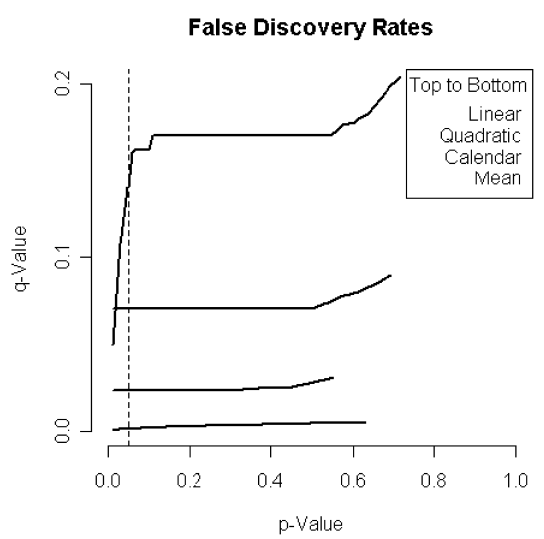


Figure 2. $q$ values versus $p$ values of tests for spline model versus four alternative models of time variation in skill.

We also apply the FDR analysis to these same bootstrap tests. As illustrated in Figure 2, the $q$ value associated with the significant $p$ values in the test of the spline model against the mean (71, or 28% of the players) is less than .02. For the test of the spline model against the mean per calendar year, the $q$ values among the significant $p$ values (13, or 5% of the players), are no more than .024. In testing the spline model against a quadratic time trend (10, or 4% of the players), the $q$ values are .071 in all cases. In testing the spline model against a linear time trend, 18 of the 25 significant $p$ values have a $q$ value of .050 with the $q$ values for the remaining seven falling between .099 and .140. On the basis of these tests, it is clear that the spline model is superior to the mean for a large number of players. Although the proportions of significant $p$ values for the other three tests are relatively low (.04 to .10), the $q$ values within these groups are also very low, indicating that almost all $p$ values less than .05 represent true rejections of the simpler models rather than false discoveries.

The overall strength of the spline model can be seen visually in Figure 1, in which very few $p$ values computed in connection with any of the four model comparisons exceed .5. (By con-

struction, when an individual $p$ value exceeds .5, the alternative model is favored in 20 or more of 40 bootstrap samples.) Under the null hypothesis $\hat{\zeta} = 0$, the distribution of $p$ values should be uniformly distributed over the [0, 1] interval. Therefore, if $\hat{\zeta} = 0$ for a given test, approximately one-half of all bootstrap $p$ values should exceed .5. The actual number of players for which the estimated $p$ value exceeds .5 ranges between 3 and 48 over the four tests. In all four model comparisons, the probability of observing values this small or lower when $\hat{\zeta} = 0$ is less than $10^{-24}$.

## 4. PLAYER SKILL

### 4.1 Time Dependence in Player Skill

The results of the preceding tests provide strong evidence that the skill levels of PGA Tour players change through time. For many players, the relationship between the player's mean skill level and time is well approximated by a linear time trend, after adjusting for autocorrelation in residual errors. For others, the relationship between mean player skill and time is more complex and cannot easily be modeled by a simple parametric-based time relationship.

Examples of time dependence in skill are illustrated in Figure 3, which shows spline fits for three selected players. All plots in the upper panel show 18-hole scores reduced by the round–course and player–course effects (connected by jagged lines) along with corresponding spline fits (smooth lines). The same spline fits (smooth lines) are shown in the lower panel along with predicted scores adjusted for round–course and player–course effects (connected by jagged lines), computed as a function of prior residual errors $\boldsymbol{\theta}$ and the first-order autocorrelation coefficient $\boldsymbol{\phi}$ estimated in connection with the spline fit.

The two plots for Chris Smith reflect 10.73 degrees of freedom used in connection with the estimate of his spline function, the largest among the 253 players. The fit for Ian Leggatt reflects $\phi = -.28$, the most negative first-order autocorrelation coefficient estimated among the 253 golfers. The standard deviation of the $\boldsymbol{\theta}$ residual errors around the spline fit for Leggatt is 2.14 strokes per round, also the lowest among the 253 players. The fit for Tiger Woods, generally regarded as the world's best player, reflects 4.11 degrees of freedom and $\phi = .08$. The resulting $u$ shape associated with Woods' fit reflects his phenomenal play in year 2000 during which he won three of the four major golf championships (The Masters, U.S. Open, British Open, and PGA Championship).

Figure 4 provides a histogram showing the distribution of degrees of freedom among the 253 spline fits. One hundred five of the fits have exactly 2 degrees of freedom, an exact linear fit, and the degrees of freedom for 175, or 69% of the splines, is less than 3. [It should be noted that, for each spline, an additional degree of freedom, not accounted for in the histograms, is used up

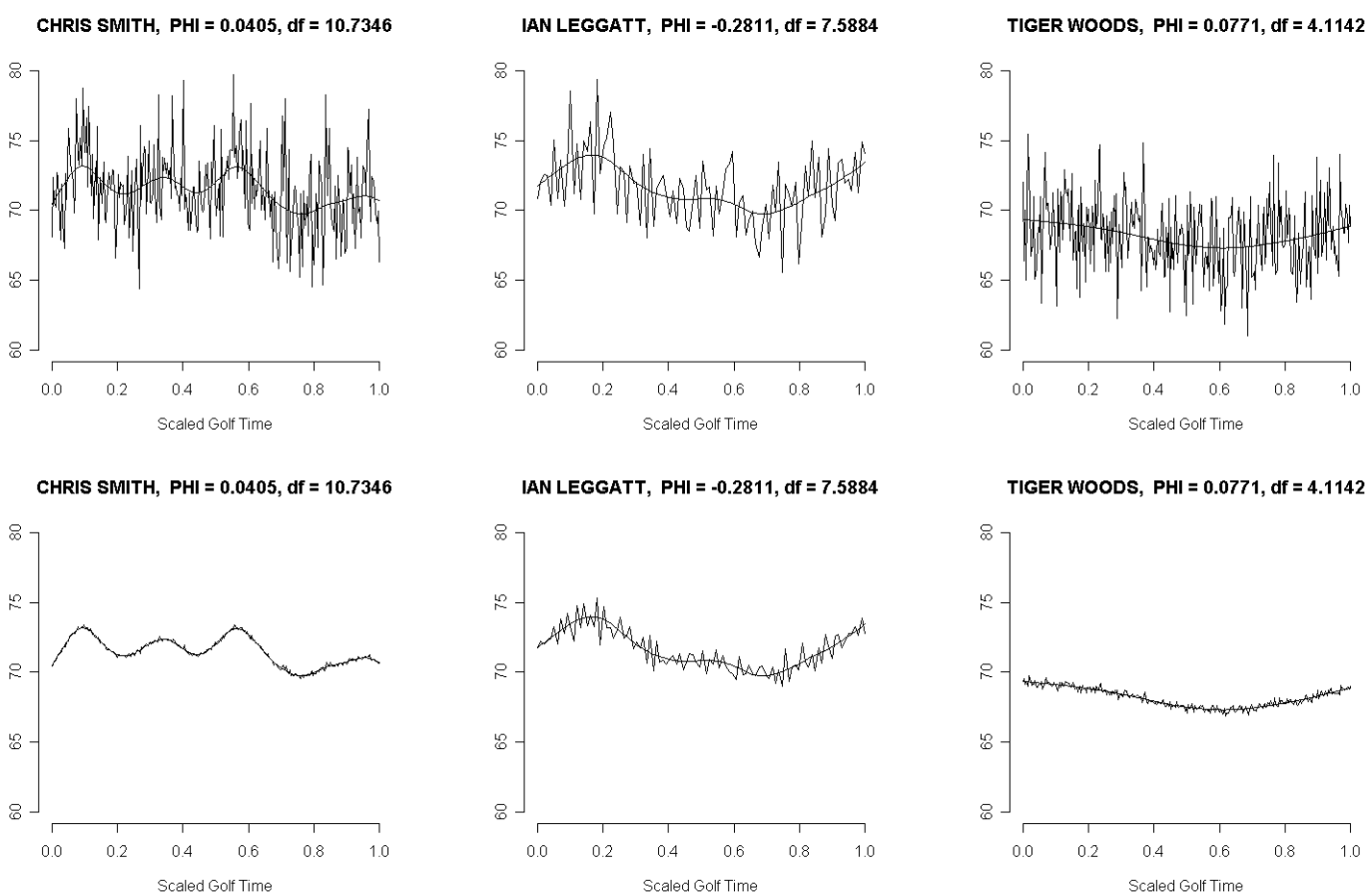


Figure 3. Spline-based mean scores adjusted for round–course effects as a function of scaled golf time. Plots in the upper panel show 18-hole scores reduced by round–course effects (connected by jagged lines) along with corresponding spline fits (smooth lines). The same spline fits (smooth lines) are shown in the lower panel along with predicted 18-hole scores adjusted for round–course effects (connected by jagged lines), computed as a function of prior residual errors and the first-order autocorrelation coefficient estimated in connection with the spline fit.

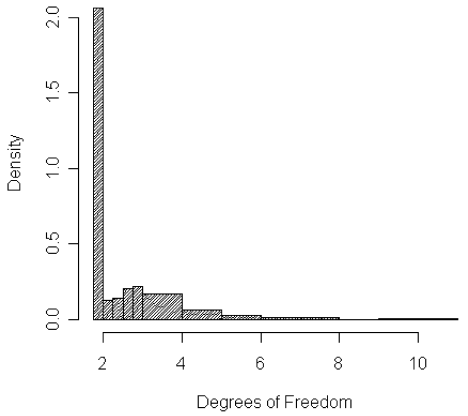


Figure 4. Degrees of freedom of cubic spline fits.

in estimating the AR(1) correlation structure of the residuals.] Seventeen of the splines are fitted with 5 or more degrees of freedom, with the maximum being 10.73 for Chris Smith.

### 4.2 Improvement and Deterioration in Skill

Figure 5 is a histogram that summarizes improvement and deterioration in skill over the 1998–2001 sample period based on changes from beginning to end in spline-based estimates of player scores adjusted for round–course and player–course effects. As illustrated in the histogram, the skill levels of 55% of the players changed by no more than one stroke per round. The skill levels of 23% of the players improved by more than one stroke, and the remaining 22% experienced skill declines by more than one stroke.

Table 1 provides a list of the 10 players showing the most improvement and deterioration in skill over the 1998–2001 sample period based on changes from beginning to end in spline-based estimates of player scores adjusted for round–course and player–course effects. Cameron Beckman was the most improved player, improving by 3.37 strokes from the beginning of 1998 to the end of 2001. Chris DiMarco and Mike Weir, relatively unknown in 1998 but now recognized among the world's elite players, were the fourth and sixth most improved players over the sample period. Bernhard Langer's improvement of 2.35 strokes per round, which led to his re-emergence among the world's top players, is reflected in the table. Among the

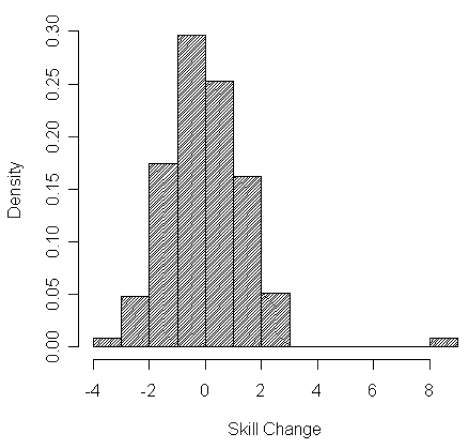


Figure 5. Changes in player skill levels.

Table 1. Players showing the most improvement and deterioration in skill over the 1998–2001 sample period

| Improving skill | | Deteriorating skill | |
|---|---|---|---|
| Player | Skill change | Player | Skill change |
| Cameron Beckman | −3.37 | Steve Jurgensen | 2.61 |
| Brian Gay | −3.17 | Tom Watson | 2.67 |
| J. J. Henry | −2.91 | Craig Stadler | 2.67 |
| Chris Dimarco | −2.86 | Craig A. Spence | 2.70 |
| Chip Beck | −2.77 | Bobby Wadkins | 2.76 |
| Mike Weir | −2.63 | Larry Rinker | 2.77 |
| Briny Baird | −2.57 | Rick Fehr | 2.77 |
| David Peoples | −2.51 | Keith Fergus | 2.85 |
| Darren Clarke | −2.39 | Lanny Wadkins | 8.20 |
| Bernhard Langer | −2.35 | Billy Ray Brown | 8.94 |

NOTE: Skill change is the difference in the spline-based estimate of a player's mean skill level from the beginning to the end of the player's golf time. Golf time is specific to each player and does not necessarily begin at the start of 1998 and end at the end of 2001.

10 golfers whose predicted scores went up the most, Billy Ray Brown, whose estimated skill level deteriorated by 8.94 strokes, most likely never fully recovered from wrist surgery. Also included were Lanny Wadkins (8.20 strokes, born 1949), Keith Fergus (2.85 strokes, 1954), Bobby Wadkins (2.76 strokes, 1951), Craig Stadler (2.67 strokes, 1953), and Tom Watson (2.67 strokes, 1949), while Fuzzy Zoeller (2.59 strokes, 1951) and David Edwards (2.38 strokes, 1956) barely missed appearing among the bottom 10. Clearly, deterioration in player skill appears to be a function of age, with a substantial amount of deterioration occurring during a golfer's mid to late 40s. But despite the natural deterioration that occurs among older players, 132 of the 253 players in the sample actually improved from the beginning of the sample period to the end. This net improvement may reflect the substantial advances in golf technology that occurred during the 1998–2001 period.

### 4.3 Ranking of PGA Tour Players

The player-specific cubic spline functions can provide point estimates of expected scores, adjusted for round–course and player–course effects, at the end of the 1998–2001 sample period and, therefore, can be used to rank the players at the end of 2001. Table 2 provides a summary of the best players among the sample of 253 as of the end of the 2001 based on the cubic spline point estimates shown in column 1 of the table. The values in column 1 can be thought of as estimates of mean player skill at the end of the sample period. By contrast, the values in column 2 can be thought of as estimates of each player's last score as a function of his ending mean skill level *and* the correlation in random errors about prior mean skill estimates.

The Official World Golf Ranking is shown in column 3 for each player as of November 4, 2001, the ranking date that corresponds to the end of the official 2001 PGA Tour season. The World Golf Ranking is based on a player's most recent rolling two years of performance, with points awarded based on position of finish in qualifying worldwide golf events on nine different tours. The ranking does not reflect actual player scores. Details of the ranking methodology can be found in the "about" section of the Official World Golf Ranking website, *www.officialworldgolfranking.com*.

Table 2. Best 20 players at the end of the 2001 season based on spline estimates of 18-hole scores adjusted for round–course and player–course effects

| Player | Spline-based estimate of score as of 11/01 | | Official World Golf Ranking as of 11/01 |
|---|---|---|---|
| | Not adjusted for autocorrelation in $\boldsymbol{\theta}$ residuals | Adjusted for autocorrelation in $\boldsymbol{\theta}$ residuals | |
| | (1) | (2) | (3) |
| 1. Bob Estes | 67.83 | 67.82 | 18 |
| 2. Davis Love III | 68.45 | 68.50 | 5 |
| 3. Phil Mickelson | 68.47 | 68.40 | 2 |
| 4. Sergio Garcia | 68.48 | 68.29 | 6 |
| 5. Dudley Hart | 68.74 | 68.55 | 49 |
| 6. Chris Dimarco | 68.84 | 68.88 | 19 |
| 7. Vijay Singh | 68.85 | 68.90 | 8 |
| 8. Tiger Woods | 68.89 | 69.00 | 1 |
| 9. Bernhard Langer | 68.93 | 69.05 | 14 |
| 10. Scott Verplank | 69.09 | 69.09 | 13 |
| 11. Kenny Perry | 69.09 | 69.24 | 34 |
| 12. Charles Howell III | 69.15 | 69.16 | 45 |
| 13. Chris Riley | 69.17 | 69.29 | 96 |
| 14. Scott McCarron | 69.20 | 69.03 | 47 |
| 15. Paul Azinger | 69.24 | 69.28 | 23 |
| 16. Darren Clarke | 69.26 | 69.08 | 10 |
| 17. David Toms | 69.28 | 69.10 | 7 |
| 18. Nick Price | 69.33 | 69.35 | 32 |
| 19. David Duval | 69.35 | 69.10 | 3 |
| 20. Ernie Els | 69.42 | 69.44 | 4 |

Perhaps the most interesting observation from Table 2 is that the spline-based estimate of Tiger Woods skill level as of November 4, 2001, adjusted and unadjusted for autocorrelation in $\boldsymbol{\theta}$ residuals, was only eighth best among the 253 players in our sample. By contrast, Woods was ranked #1 according to the Official World Golf Ranking. The spline-based estimate of Woods' skill level reflects his widely publicized slump during the second half of 2001. Woods won 4 of the first 10 tournaments of 2001, placing 5th, on average, during this period. During the final nine tournaments of 2001, Woods won only one tournament and averaged a 17th-place finish. In contrast to the spline-based estimate of Woods' skill level, the Official World Golf Ranking, based on two years of prior performance, failed to pick up Tiger's slump.

Among the 20 players listed in Table 2, the predicted score for Bob Estes, adjusted and unadjusted for autocorrelation in residual errors (columns 1 and 2), is the lowest. Although it may come as a surprise that a player with as little name recognition as Estes would be predicted to shoot the lowest score at the end of 2001, Estes exhibited marked improvement over the last quarter of the sample period, and the improvement was sufficiently pronounced to be reflected in his cubic spline–based skill estimate.

## 5. SOURCES OF VARIABILITY IN GOLFER PERFORMANCE

### 5.1 Decomposing Variability

By combining (1) and (2) and rearranging, the sources of random variation in player scores can be expressed as $\mathbf{s} - \mathbf{P}h(\bullet) = \mathbf{P}\boldsymbol{\theta} + \mathbf{R}b_2 + \mathbf{C}b_3$, indicating that unusual performance may be due to any of three factors: player-specific effects, round–course effects, and player–course effects. To what extent, if any, can we attribute these three sources of variation to luck?

We believe that professional golfers think of luck as sources of variation in scoring outside a player's direct and conscious control. For example, if a player is assigned a relatively easy course rotation in a multiple-course tournament, professionals would say this player had good luck in his course assignments. We can estimate the extent of such "luck" through the round–course effect. Similarly, if a tournament happens to be played on a course that favors a particular player's style, players might attribute any favorable outcome associated with playing on this particular course to luck, because a player cannot choose the course on which a tournament is played. We estimate the magnitude of this "luck" through player–course effects. Any remaining variation in score, not attributable to round–course and player–course effects, is reflected in the $\boldsymbol{\theta}$ error. Clearly, some of the $\boldsymbol{\theta}$ error reflects variation due to easily recognizable factors that we do not measure, that is, lucky bounces, good and bad lies, relatively favorable or unfavorable weather conditions, and so on. But some may not be nearly as easy for observers to identify and may simply represent physical variation in a player's swing. For example, given the intrinsic skill level of a particular player, assume that there is a 50% chance that he will hit the green from a fairway shot 220 yards out. If he hits five such greens in a row, and his intrinsic skill level has not changed, we would say this player experienced good luck (favorable variation). The root cause may be favorable variation in his swing, but if the player cannot maintain sufficient control over his swing to sustain this favorable variation, we would call it luck when he hits five greens in a row. On the other hand, if he can sustain the favorable variation, this should lead to a change in skill.

In this section we discuss player-specific sources of variation ($\boldsymbol{\theta}$), round–course effects, and player–course effects and summarize their relative importance in determining outcomes in PGA Tour events. In Section 6 we study autocorrelation in player-specific error ($\boldsymbol{\theta}$).

### 5.2 Player–Course Effects

The absolute magnitude of random player–course effects is quite small, ranging from $-.065$ to .044 over all possible 12,485 nested player–course interactions. If the model is fit, imposing the restriction that the first-order autocorrelation coefficient $\phi_i$ is 0 for all players, the range in player–course effects is $-.21$ to .14. As discussed in the next section, the autocorrelation in $\boldsymbol{\theta}$ residuals is positive for 61% of the players. If a player's $\boldsymbol{\theta}$ residuals are positively autocorrelated, good (bad) rounds would tend to be followed by more good (bad) rounds. Therefore, failing to account for autocorrelation when all or most of the rounds in a tournament are played on the same course may cause the absolute magnitude of player–course effects to appear larger than they actually are. We note, in passing, that even though these effects are small, we find that including them in the model produces substantially more curvature in the spline fits.

### 5.3 Round–Course Effects

Figure 6 summarizes the distribution of 848 random round–course effects estimated in connection with our model. The effects range in value from −3.92 to 6.95, implying almost an 11-stroke difference between the relative difficulty of the most difficult and easiest rounds played on the Tour during the 1998–2001 period.

Over this period, 25 tournaments were played on more than one course. In multiple-course tournaments, players are (more or less) randomly assigned to a group that rotates among all courses used for the tournament. By the time each rotation is completed, all players will have played the same courses. At that time, a cut is made, and participants who survive the cut finish the tournament on the same course. Although every attempt is made to set up the courses so that they play with approximately the same level of difficulty in each round, tournament officials cannot control the weather and, therefore, there is no guarantee that the rotation assignments will all play with the same levels of difficulty.

Figure 7 shows the distribution of the difference in the sum of round–course effects for the easiest and most difficult rotations for the 25 tournaments that were played on more than one course. To illustrate the values shown in this figure, suppose half the players in a tournament play course A on day 1 and then play course B on day 2, and the other half plays B and then A. On day 1, the round–course effects for courses A and B are 3.0 and 1.7, respectively, and on day 2, the round–course effects are 6.0 and 2.2. Then the rotation B–A will have played $(1.7 + 6.0) - (3.0 + 2.2) = 2.5$ strokes more difficult than the A–B rotation. For seven of 25 tournaments, the difference in difficulty between the easiest and most difficult rotations was less than .50 strokes. Thus, on average, the difference for these tournaments was sufficiently small that a player's total score should have been the same regardless of the course rotation to which he was assigned. At the extreme, there was a 5.45- and 5.26-stroke differential between the relative difficulties of the easiest and two most difficult rotation assignments in the 1999 ATT Pebble Beach Pro-Am. The top nine finishers in this tournament all played an easy rotation, as did four of the five players who tied for 10th place. Among the top 20 finishers, only two played one of the two difficult rotations. Clearly, for this particular tournament, variability in course assignments had much to do with determining the top finishers. However, for the 157 tournaments in our sample played on the same course, round–course effects have no impact in determining tournament outcomes, because all golfers play the same course each day.

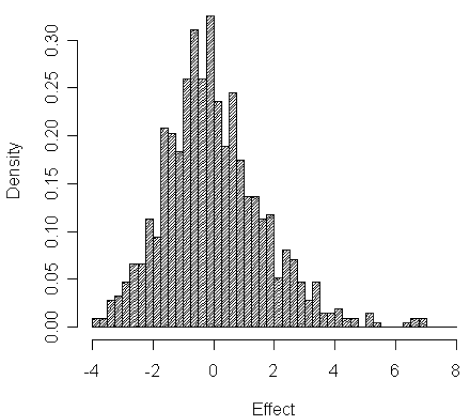


Figure 6. Distribution of round–course effects.

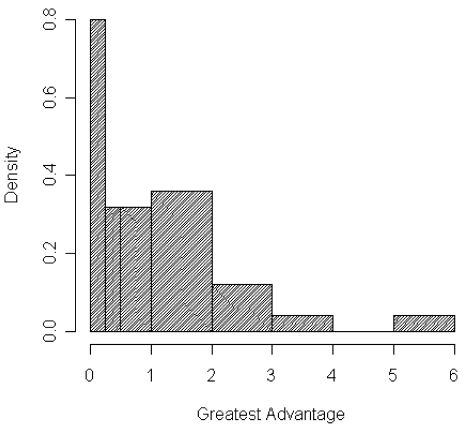


Figure 7. Distribution of greatest advantage in multiple-course tournaments.

### 5.4 $\theta$ Residuals

Figure 8 provides a histogram summarizing the standard deviation of $\boldsymbol{\theta}$ residual errors among all 253 players in the sample. The range of standard deviations is 2.14–3.44 strokes per round, with a median of 2.68 strokes. John Daly and Phil Mickelson, both well known for their aggressive play and propensities to take risks, have the 3rd and 13th highest standard deviations, respectively. Ian Leggatt has the lowest standard deviation. Chris Riley and Jeff Sluman, both known as very conservative players, have the second and fifth lowest deviations, respectively.

It is interesting to consider whether average scores and standard deviations of $\boldsymbol{\theta}$ residual errors are correlated. A least squares regression of standard deviations against the mean of each player's spline-based estimate of skill over the entire sample period yields Expected score $= 66.191 + 1.91 \times$ Standard deviation, with adjusted $R^2 = .056$, $F = 15.9$, and $p$ value $< .0001$. Thus, there is a tendency for greater variation in player scores to lead to slightly higher average scores.

### 5.5 What Does It Take to Win a PGA Tour Event?

It is interesting to consider the relative contributions of each source of variability to winning or finishing among the top play-

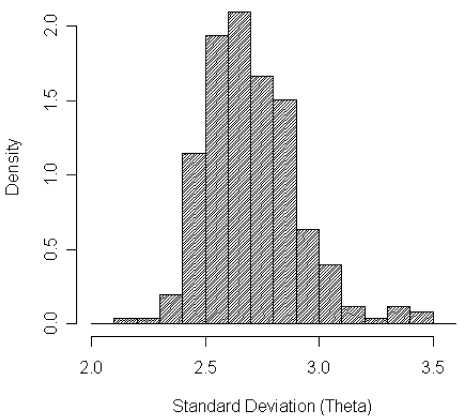


Figure 8. Distribution of standard deviations of theta residuals.

Table 3. Actual scores and residual scores for top finishers in 2001 Players Championship and 2000 Byron Nelson Classic

| | Four-round totals | | | | |
|---|---|---|---|---|---|
| | Score | $\theta$ residual | Round–course effect | Player–course effect | Total residual |
| *2001 Players Championship* | | | | | |
| Tiger Woods | 274 | −8.623 | 0 | −.040 | −8.662 |
| Vijay Singh | 275 | −12.049 | 0 | −.056 | −12.105 |
| Bernhard Langer | 276 | −13.166 | 0 | −.080 | −13.247 |
| Jerry Kelly | 278 | −14.667 | 0 | −.079 | −14.745 |
| Billy Mayfair | 281 | −11.417 | 0 | −.107 | −11.524 |
| Hal Sutton | 281 | −9.552 | 0 | −.207 | −9.759 |
| Frank Lickliter | 282 | −9.982 | 0 | −.083 | −10.065 |
| Paul Azinger | 282 | −7.095 | 0 | .035 | −7.060 |
| Scott Hoch | 282 | −7.817 | 0 | −.170 | −7.988 |
| Joe Durant | 283 | −8.307 | 0 | .016 | −8.290 |
| Nick Price | 283 | −5.448 | 0 | −.170 | −5.618 |
| David Toms | 284 | −5.107 | 0 | −.062 | −5.169 |
| Jose Maria Olazabal | 284 | −8.404 | 0 | −.032 | −8.436 |
| Tom Lehman | 284 | −5.499 | 0 | −.133 | −5.632 |
| Franklin Langham | 285 | −11.356 | 0 | −.025 | −11.381 |
| Scott Dunlap | 285 | −8.815 | 0 | −.140 | −8.955 |
| Jonathan Kaye | 286 | −7.128 | 0 | −.021 | −7.149 |
| Kenny Perry | 286 | −3.758 | 0 | .025 | −3.733 |
| Lee Janzen | 286 | −6.629 | 0 | −.125 | −6.754 |
| Corey Pavin | 287 | −7.475 | 0 | −.020 | −7.495 |
| J. P. Hayes | 287 | −7.786 | 0 | −.027 | −7.813 |
| Jim Furyk | 287 | −2.484 | 0 | .035 | −2.449 |
| Robert Allenby | 287 | −5.003 | 0 | −.036 | −5.039 |
| Tim Herron | 287 | −6.181 | 0 | .037 | −6.144 |
| *2000 Byron Nelson Classic* | | | | | |
| Davis Love III | 269 | −8.892 | 1.147 | −.026 | −7.771 |
| Jesper Parnevik | 269 | −8.122 | −1.147 | −.011 | −9.280 |
| Phil Mickelson | 269 | −8.024 | 1.147 | −.028 | −6.905 |
| John Huston | 270 | −9.305 | −1.147 | −.040 | −10.492 |
| Tiger Woods | 270 | .672 | 1.147 | .046 | 1.865 |
| Bob Estes | 271 | −11.458 | 1.147 | −.033 | −10.344 |
| Brandel Chamblee | 272 | −12.280 | 1.147 | −.001 | −11.134 |
| Mark Brooks | 272 | −9.879 | −1.147 | −.051 | −11.078 |
| Jerry Smith | 273 | −10.757 | −1.147 | −.033 | −11.938 |
| Paul Stankowski | 274 | −10.307 | 1.147 | −.060 | −9.220 |
| Scott Dunlap | 274 | −7.913 | 1.147 | −.009 | −6.775 |
| Tommy Armour III | 274 | −9.449 | −1.147 | −.078 | −10.674 |
| Glen Day | 275 | −5.563 | −1.147 | −.016 | −6.726 |
| Andrew Magee | 276 | −6.152 | −1.147 | −.051 | −7.350 |
| Ben Bates | 276 | −9.398 | −1.147 | −.067 | −10.612 |
| Blaine McCallister | 276 | −6.761 | −1.147 | −.032 | −7.940 |
| Nick Price | 276 | .103 | −1.147 | −.005 | −1.049 |
| Sergio Garcia | 276 | −1.238 | −1.147 | −.032 | −2.418 |
| David Duval | 277 | .195 | 1.147 | −.006 | 1.336 |
| David Toms | 277 | −2.979 | 1.147 | −.004 | −1.836 |
| Franklin Langham | 277 | −2.696 | −1.147 | −.001 | −3.844 |
| Jim Furyk | 277 | .838 | −1.147 | −.001 | −.311 |
| Shigeki Maruyama | 277 | −3.350 | −1.147 | −.023 | −4.521 |

NOTE: Total residual is the sum of the $\theta$ residual, the round–course effect, and the player–course effect. The round–course effect for each tournament round is computed by taking the difference between the round–course effect associated with a player's score and the average of all round–course effects for the same day. These differences are then summed over the four tournament rounds.

ers in a PGA Tour event. Table 3 summarizes the actual final tournament scores and the four-round total of $\theta$ residuals, round–course effects, and player–course effects for the top-20 finishers and ties in the 2001 Players Championship and the 2000 Byron Nelson Classic. The entire Players Championship was played on a single course, and, therefore, there was no dif-

ference in the round–course effects among the participants. By contrast, the first two rounds of the Byron Nelson Classic were played on two different courses with the same type of rotation as described in Section 5.3.

The patterns of residuals exhibited in the table are typical of those for all 182 tournaments. Among the top-20 finishers in the 2001 Players Championship, the total residual score of all players is negative. The round–course effect is 0 in all cases, and the player–course effect is too small to have affected the order of finish. Among the top-20 finishers in the Byron Nelson Classic, all but Tiger Woods and David Duval experienced negative total residuals. The total $\boldsymbol{\theta}$ residual for both of these players was close to 0, indicating that they could have finished among the top 20 by playing at their average skill levels. It is interesting to note that the Byron Nelson tournament was won by Jesper Parnevik in a three-way playoff with Davis Love III and Phil Mickelson. All three players shot 269 during regular play, and all three had similar $\boldsymbol{\theta}$ residuals, indicating that each played approximately the same relative to his skill level. However, Parnevik had a $1.147 - (-1.147) = 2.294$ stroke advantage over the other two due to a favorable course rotation assignment during the first two rounds. Therefore, based on our model, we conclude that Parnevik's presence in the playoff was partially due to the luck of the rotation assignment.

We note also that to have won these tournaments, and almost all others in the sample, not only must one have played better than normal, but one must have also played sufficiently well (or with sufficient luck) to overcome the collective good luck of many other participants in the same event. Specifically, over all 182 tournaments, the average total $\boldsymbol{\theta}$ residual score for winners and first-place ties was $-9.64$ strokes, with the total $\boldsymbol{\theta}$ residual ranging from $+.13$ strokes for Tiger Woods in the 1999 Walt Disney World Resort Classic to $-21.59$ strokes for Mark Calcavecchia in the 2001 Phoenix Open. Table 4 summarizes the highest 20 total residual scores per tournament for winners and first-place ties. It is noteworthy that only one tournament was won by a player with a positive total residual score, the 1999 Disney Tournament won by Tiger Woods. It is also noteworthy that Woods' name appears 13 times in Table 4 and that all but one of the other players on the list (Phil Mickelson, Sergio Garcia, David Duval, and Davis Love III) are among the world's best players. Thus, over the 1998–2001 period, only Tiger Woods, and perhaps a handful of other top players, were able to win tournaments without experiencing exceptionally good luck.

## 6. PERSISTENCE IN RESIDUAL PERFORMANCE

As noted in Section 4.1, and reflected in the shapes of the 253 individual spline fits, our tests provide strong evidence that the skill levels of PGA Tour players change through time. These tests show the spline model to be a superior specification relative to four alternatives, especially the specification that assumes a constant level of skill for each player. In this section we focus on persistence in player performance after accounting for skill changes, specifically testing for first-order autocorrelation in $\boldsymbol{\theta}$ residuals. We also discuss the connection between this analysis and the hot-hands tests in the sports statistics literature.

### 6.1 Autocorrelation in $\boldsymbol{\theta}$ Residuals

In this section, we describe three sets of tests for positive and negative first-order autocorrelation in $\boldsymbol{\theta}$ residuals. The first set employs the bootstrap and is based on the same bootstrap sample described in connection with tests of the spline model against four alternative model specifications and uses the same

Table 4. Twenty highest total residual scores of tournament winners or players who tied for first

| Tournament | Winning score | Winner or tie for first | Total $\boldsymbol{\theta}$ residual | Total round–course effect | Total player–course effect | Total residual |
|---|---|---|---|---|---|---|
| 99 Disney | 271 | Tiger Woods | .13 | .10 | .01 | .24 |
| 99 Tour Champ | 269 | Tiger Woods | −.39 | .00 | .00 | −.39 |
| 99 AM Express | 278 | Tiger Woods | −1.32 | .00 | .01 | −1.30 |
| 00 ATT | 273 | Tiger Woods | −.42 | −1.17 | −.01 | −1.60 |
| 00 Canadian Open | 266 | Tiger Woods | −2.64 | .00 | −.02 | −2.66 |
| 99 NEC | 270 | Tiger Woods | −2.57 | .00 | −.12 | −2.70 |
| 99 Western | 273 | Tiger Woods | −3.59 | .00 | .07 | −3.51 |
| 00 Mercedes | 276 | Tiger Woods | −3.54 | .00 | .02 | −3.52 |
| 01 Tour Championship | 270 | Sergio Garcia | −3.50 | .00 | −.03 | −3.53 |
| 01 Disney | 266 | Davis Love III | −3.72 | .23 | −.05 | −3.54 |
| 00 Bell South | 205 | Phil Mickelson | −3.58 | .00 | −.04 | −3.62 |
| 99 PGA | 277 | Tiger Woods | −3.89 | .00 | −.03 | −3.92 |
| 98 Houston | 276 | David Duval | −4.01 | .00 | .07 | −3.94 |
| 01 Buick Invitational | 269 | Phil Mickelson | −3.25 | −.79 | −.03 | −4.07 |
| 00 Bay Hill | 270 | Tiger Woods | −4.29 | .00 | .01 | −4.27 |
| 00 PGA | 270 | Tiger Woods | −4.46 | .00 | −.03 | −4.49 |
| 01 Bay Hill | 273 | Tiger Woods | −4.68 | .00 | .01 | −4.66 |
| 00 British | 269 | Tiger Woods | −4.67 | .00 | −.03 | −4.70 |
| 98 Texas Open | 270 | Hal Sutton | −4.91 | .00 | .00 | −4.91 |
| 01 Hartford | 264 | Phil Mickelson | −4.93 | .00 | −.03 | −4.96 |

NOTE: Total residual is the sum of the $\boldsymbol{\theta}$ residual, the round–course effect, and the player–course effect. The round–course effect for each tournament round is computed by taking the difference between the round–course effect associated with a player's score and the average of all round–course effects for the same day. These differences are then summed over the tournament rounds.

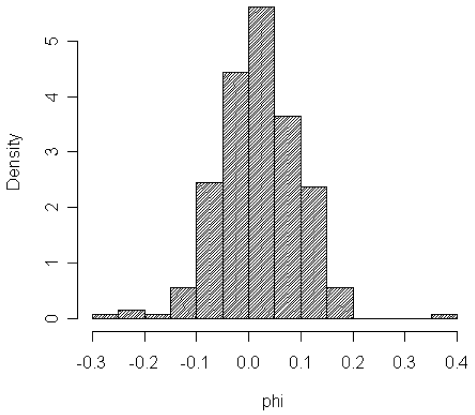


Figure 9. Distribution of first-order autocorrelation coefficients.

interpolation method in computing $p$ values (see Sec. 3.5). As illustrated in Figure 9, the residual $\boldsymbol{\theta}$ errors of 155 players (61%) are positively correlated.

In the first set of tests, we find that 12 of 253 autocorrelation coefficients are significantly negative at the 5% level and 24 are significantly positive. In a two-tail test, 23 coefficients are significantly different from 0. We also apply Storey's FDR analysis to the $p$ values associated with tests for positive and negative autocorrelation and find that the $q$ value for 3 of the 12 significantly negative coefficients is .68. For the remaining 9, the $q$ value falls between .78 and .83, indicating that a large portion of these 9 coefficients are actually 0. Among the 24 significantly positive coefficients, 15 are associated with $q$ values of .18. The range of $q$ values for the remaining 9 is .32 to .43. The histogram of $p$ values in Figure 10 suggests that $\phi = 0$ for a large portion of players. However, consistent with the analysis of $q$ values summarized previously, the large number of $p$ values in the range of 0–.10 suggests that $\phi > 0$ for some players.

Using the same bootstrap samples, we also conduct two additional sets of bootstrap tests for negative and positive autocorrelation. In both tests, we employ Fisher's transformation of the correlation coefficients to produce values closer to normality, although this transformation makes little difference in any of our results, because most of the coefficients are close to 0. In the first test, sometimes referred to as a parametric bootstrap, we assume that each player's set of transformed autocorrelation coefficients is normally distributed with a standard deviation equal to the standard error computed from the player's 40 Fisher-transformed coefficients. Under the normality assumption, the number of significantly negative and positive autocorrelation coefficients is 7 and 23, respectively. Among the seven significantly negative coefficients, the $q$ value is .33 for two of the seven and .999 for the remaining five. Among the 23 significantly positive coefficients, one is associated with a $q$ value of .02, whereas the $q$ values for the remaining 22 range from .14 to .28.

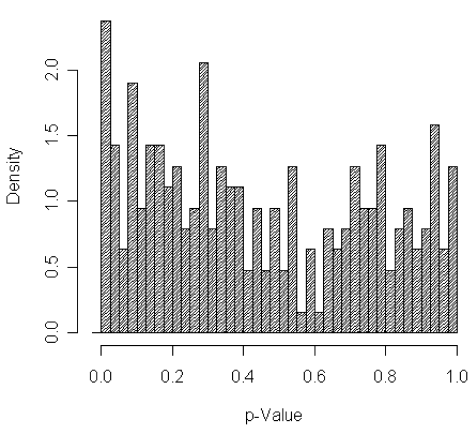


Figure 10. Distribution of $p$ values in bootstrap test for first-order autocorrelation.

In our final test, we compute $z$ values of Fisher-transformed autocorrelation coefficients for each player, aggregate these $z$ values across all 253 players, and assume that the cumulative probabilities from the aggregate distribution apply to each player. Using this approach, we are able to employ the information from all 10,120 bootstrap samples in drawing inferences for each player rather than having to use only 40 bootstrap samples per player. Among seven significantly negative coefficients, the $q$ value is .06 for two and 1.00 for the remaining five. Among the 23 players with significantly positive coefficients, the same 23 players as in the previous test, the $q$ values for five fall approximately uniformly between 0 and .17, with the remaining 18 $q$ values ranging from .18 to .34.

Only two players show any evidence of significant negative autocorrelation in combination with low $q$ values, and this evidence only occurs in our third test. Therefore, we conclude that there may be evidence of significant negative autocorrelation for two players at most. Twenty players show evidence of significantly positive autocorrelation in all three tests, and 23 show evidence of significant positive autocorrelation in the second two tests. Of these 23 golfers, the maximum $q$ value is .37. On the basis of both $p$ values and $q$ values, we estimate that approximately $100 - 37 = 63\%$ of the 23 significantly positive coefficients are indeed significantly positive.

## 6.2 Streaky Play

Streaky play is the tendency for abnormally good and poor performance to persist over time. "Hot hands," generally thought of as streaky play over very short intervals of activity, has been the focus of a number of statistical studies of sports. Although much of the hot-hands literature focuses on basketball and baseball, there is some literature related to streaky play in golf. Gilden and Wilson (1995) studied hot/cold hands in putting and found more subjects displayed streaks in putting success/failure than would be expected by chance alone (using runs tests on success versus failure). Further, they found that putting streaks were more likely when the difficulty of the task was commensurate with the golfer's skills: Streaks were much less likely when the task was very difficult relative to skill.

Clark studied streakiness in 18-hole golf scores (2003a, b, 2004a) and also in hole-by-hole scores (2004b). (An excellent summary of the psychological dimensions of streaky play in golf is provided in Clark 2003a.) In general, Clark found little evidence for streaky play among participants on the PGA Tour, Senior PGA (SPGA) Tour, and Ladies PGA (LPGA) Tour. Throughout his studies, Clark defined a success as a score that equaled or exceeded par, either on an 18-hole or a hole-by-hole basis. Although he found some evidence for streaky play in 18-hole scores, he showed that this evidence was the result of par

or better being a more frequent occurrence on the easiest golf courses. In a related set of articles, Clark (2002a, b) found no evidence of choking among players who were in contention going into the final rounds of regular PGA, SPGA, and LPGA Tour events and the PGA Tour's annual qualifying tournament known as Q–School. In none of these articles was player performance measured relative to a player's normal level of play or in relation to the relative difficulty of each round. Therefore, Clark's results are not directly comparable to ours.

Compared with other competitive sports, measuring success and failure in golf can be done with greater precision. Unlike competitors in team sports, golfers engaged in stroke play do not have to face defenses or deal with the responses of teammates who may attempt to adapt to their level of play. Although each stroke in golf represents a unique challenge, the strokes taken collectively over an entire 18-hole round represent a reasonably homogeneous challenge for most competitors in a golfing event. And unlike a sport such as basketball, in which a player either makes or misses a basket, it is much easier to measure varying degrees of success and failure in golf. Thus, conventional tests of hot hands, which tend to treat success and failure in terms of binary outcomes, may not be appropriate for golf.

We believe that the most compelling tests for streakiness in golf, where performance is measured at the round (18-hole) level, are our previously documented bootstrap tests of the significance of the first-order autocorrelation coefficients associated with the $\boldsymbol{\theta}$ residuals estimated in connection with each of our spline fits. Because our spline fits should account for any changes in skill over time, significant positive autocorrelation in the residual errors is appropriately interpreted as evidence of streaky play. Based on three sets of bootstrap tests, we conclude that the estimated autocorrelation coefficient is significantly positive for as many as 23 golfers. Thus, there is clearly a tendency for some PGA Tour participants to experience streaky play above and beyond any changes that might also be taking place in their estimated mean skill levels.

### 6.3 Applying More Conventional Tests of Hot Hands to Golf

In addition to noting the frequency of significant positive first-order autocorrelation in $\boldsymbol{\theta}$ residuals, we apply two other tests for streaky play that have been used in the hot-hands sports literature. The first is a standard runs test, and the second, used by Albright (1993) to test hitting streaks in baseball, tests whether a string of successive residual golf scores follows a first-order Markov chain. Both tests are conducted on $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ residual scores. Due to the degree of autocorrelation in the $\boldsymbol{\theta}$ residuals, we expect to find confirming evidence of streaky play from the additional tests as applied to these residuals. However, we do not expect to find evidence of streaky play when applying these tests to the $\boldsymbol{\eta}$ residuals (where the effects of first-order autocorrelation have been removed). Also, each of these tests is performed on residual scores transformed into a $\{1, 0\}$ binary form representing golf scores that are either lower (success) or higher (failure) than predicted by our model. Although binary form is the only practical form for measuring success or failure when shooting basketballs or hitting baseballs, much of the information about the degree of success or failure in golf is lost when residual golf scores are transformed in this fashion. As a result, the power of these tests applied to binary golf scores should be somewhat lower than our bootstrap and parametric bootstrap tests for significant first-order autocorrelation in individual residual scores.

In both tests, we define a player's expected residual score, given the residuals of his most recent run, as

$$\text{Expected residual} = \frac{-\text{Sum of residuals in most recent run}}{n - \text{Number of residuals in most recent run}},$$

because the total of his $n$ residual scores must sum to 0. Following Mendenhall, Scheafler, and Wackerly (1981, p. 601), when runs tests are applied to $\boldsymbol{\theta}$ errors, 22 of 253 $z$ statistics are significant at the 5% level. Ten of these 22 players are among the 23 for whom our second two bootstrap tests for positive autocorrelation are significant. The $q$ value associated with the six most significant $p$ values ranges between .02 and .10, but for the remaining 16, the range of $q$ values is .24 to .46. These results corroborate those of the bootstrap tests, suggesting that more players experienced episodes of streaky play over our sample period than can reasonably be expected by chance. On the other hand, runs tests conducted on $\boldsymbol{\eta}$ errors are significant at the 5% level for only 7 of 253 players. Among these 7, the $q$ value for each is .59, suggesting that no more than two or three truly represent departures from white noise. Overall, the runs tests appear to provide no additional evidence of streaky play beyond the evidence provided by the tests for first-order autocorrelation.

Albright (1993) studied hitting streaks in baseball using several methods. The test statistic is given by $\chi^2 \sim \mathrm{M}(M_{0,0}M_{1,1} - M_{1,0}M_{0,1})^2/M_0^2M_1^2$, where $M_{i,j}$ is the number of times that state $i$ is followed by state $j$ and $M_i$ is the number of times the $i$th state occurs. Under the null hypothesis of randomness, this statistic has a chi-squared distribution with 1 degree of freedom. In our application, the two states represent better and worse-than-average performance, as estimated by negative and positive $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ residuals from our model.

We apply this method to the 253 golfers in our sample for both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ errors. When we use the $\boldsymbol{\theta}$ residuals from the spline, 18 golfers have chi-squared test statistics that reject the null hypothesis of no dependence at the 5% level. The corresponding $q$ values for these 18 golfers fall between .085 and .433, suggesting that approximately one-third may represent false discoveries. For the $\boldsymbol{\eta}$ residuals, we find that only three golfers show evidence of nonrandom residual scores, but the $q$ values for all three are .995, indicating that each is a false discovery. As with the runs tests, these tests provide no additional evidence of streaky play beyond that provided by the tests for first-order autocorrelation.

## 7. SUMMARY AND CONCLUSIONS

In this study we develop a generalized additive model to estimate cubic spline–based time-dependent mean skill functions and the first-order autocorrelation of residual scores about the mean (using the model of Wang 1998) for 253 active PGA Tour golfers over the 1998–2001 period, while simultaneously adjusting for round–course and player–course random effects. Using this method, we are able to estimate time-varying measures

of skill and luck for each player in the sample. Estimates of spline-based player-specific skill functions provide strong evidence that the skill levels of PGA Tour players change through time. For many players, the relationship between the player's average skill level and time is well approximated by a linear time trend, but for others, the relationship between mean player skill and time is more complex and cannot easily be modeled by a simple parametric-based time relationship. Using the model, we are also able to identify the players who experienced the most improvement and deterioration in their skill levels over our sample period. Generally, the oldest players on the Tour experienced the most significant deterioration in skill. An interesting area for further research is the relationship between skill changes revealed by spline estimates, player characteristics, technology, and changes in course design.

With our model, we are able to rank golfers at the end of 2001 on the basis of their estimated mean skill levels and also provide point estimates of their end-of-2001 scores taking into account both the mean skill functions and the simultaneously estimated first-order autocorrelations in residual scores. Our rankings are generally consistent with the Official World Golf Ranking, but we are able to capture trends in recent performance, such as Tiger Woods' well-documented slump during the second half of the 2001, that are not well reflected in the Official World Golf Ranking.

We also address the role of luck in golf. In our sample, the range of standard deviations in residual golf scores is 2.14–3.45 strokes per round, with a median of 2.69 strokes. As one might expect, John Daly and Phil Mickelson, both known for their aggressive play and propensities to take risks, are among the players with the highest standard deviations. We find that there is a tendency for higher standard deviations to be associated with slightly higher average scores.

We find that mean skill alone is insufficient to win a golf tournament; a little luck (i.e., unusually favorable outcomes and/or skilled play) is required for the most highly skilled players such as Tiger Woods, and lots of luck is required for more average players to win. Although the inclusion of player–course interaction effects is important in obtaining a proper specification of mean player skill, the general magnitude of these effects is too small to have much of an effect on actual tournament outcomes. By contrast, we do find that the particular course rotations to which players are assigned can have a pronounced effect on the outcomes of tournaments played on more than one course. For 7 of 25 such tournaments, the rotations made a difference of less than .50 strokes per round, a difference sufficiently small to keep the score for a typical player unaltered. At the extreme, however, there was a 5.45-stroke differential between the relative difficulties of the easiest and most difficult course rotation assignments.

The possible existence of streaky play has been studied extensively in the statistical sports literature. In a series of articles, Clark concluded that there was little evidence of streaky play in golf. By contrast, we find that 155 of 253 first-order autocorrelation coefficients associated with residual golf scores are positive, and among these 155 coefficients, 23 are significant at the 5% level, most of which do not appear to be false discoveries. Thus, there is clearly a tendency for a small number of PGA Tour participants to experience statistically significant streaky play. We find confirming evidence of streaky play with a conventional runs tests and a Markov chain test, but these tests do not appear to provide any additional evidence of streakiness beyond that provided by the tests for first-order autocorrelation.

We believe that our approach to modeling PGA Tour golf scores can be applied to any sport for which each participant receives an objectively determined score, faces no defense from or physical or strategic interactions with other participants, and for which the physical conditions of play might vary from one round of play to the next. Examples include timed winter events such as downhill skiing, bobsled, and luge, field events such as javelin and shot put, and perhaps other sports such as bowling and shooting. We also believe that our approach to performance measurement can be applied in numerous finance and economics settings in which the mean level of performance may vary with time and require adjustments to reflect risk classes, portfolio types, industry, regulatory, and country classifications, and so on. Examples include mutual fund performance evaluation, the evaluation of portfolio selection and trading rule strategies, and the role of skill and luck in executive performance.

## APPENDIX: MODEL ESTIMATION

Our model, as expressed in the following discussion in terms of (1) and (2), represents a generalized additive model that can be estimated using backfitting as described in Hastie and Tibshirani (1990):

$$\mathbf{s} = \mathbf{P}(h(\bullet) + \boldsymbol{\varphi} + \boldsymbol{\eta}) + \mathbf{R}b_2 + \mathbf{C}b_3.$$

In the model, $h(\bullet) = (h_1(\bullet), \ldots, h_{253}(\bullet))'$, $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_{253})'$, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{253})'$, where $h_i(\bullet)$, $\boldsymbol{\varphi}_i$, $\mathbf{s}$, $\mathbf{P}$, $\mathbf{R}$, $\mathbf{C}$, $b_2$, and $b_3$ are defined in the main body of the article. The steps for estimating the model are as follows:

1. Compute initial estimates of $b_2$ and $b_3$ (details are given later in steps a–c).
2. Compute $\mathbf{s}_1 = \mathbf{s} - \mathbf{R}b_2 - \mathbf{C}b_3$.
3. Estimate $\mathbf{s}_1 = \mathbf{P}(h(\bullet) + \boldsymbol{\varphi} + \boldsymbol{\eta})$.
4. Compute $\mathbf{s}_2 = \mathbf{s} - \mathbf{P}(h(\bullet) + \boldsymbol{\varphi})$.
5. Estimate $\mathbf{s}_2 = \mathbf{R}b_2 + \mathbf{C}b_3 + \boldsymbol{\varepsilon}$.
6. Loop back to step 2 until convergence.

The time required to compute one cycle of steps 2–6 is approximately three hours using a Windows-based PC running with a 2.80-GHz Intel® Xeon™ processor. However, if it is assumed that $\phi_1 = \phi_2 = \cdots = \phi_{253} = 0$ when estimating $h(\bullet)$, the time required to cycle through 50 iterations of steps 2–6 is only 11 hours. Therefore, to the extent possible, we attempt to perform the bulk of our backfitting calculations assuming $\phi_1 = \phi_2 = \cdots = \phi_{253} = 0$. This is accomplished in step 1 in which we compute our starting estimates of $b_2$ and $b_3$ as follows:

a. Obtain initial estimates of $b_2$ and $b_3$ by estimating $\mathbf{s} = \mathbf{P}b_1 + \mathbf{R}b_2 + \mathbf{C}b_3 + \boldsymbol{\varepsilon}$.

b. Cycle through 50 iterations of steps 2–6 while assuming $\phi_1 = \phi_2 = \cdots = \phi_{253} = 0$.

c. Denote the estimates of $b_2$ and $b_3$ after 50 iterations as $b_2^0$ and $b_3^0$, respectively, the corresponding vector of spline fits as

$h^0(\bullet) = (h_1^0(\bullet), \ldots, h_{253}^0(\bullet))'$, and the corresponding vector of autocorrelated error term components as $\boldsymbol{\varphi}^0 = (\varphi_1^0, \ldots, \varphi_{253}^0)'$. Then the starting estimate of $b_2$ in step 1 is computed as

$$b_2 \approx b_2^0 + \frac{\partial b_2^0}{\partial \phi_1}\Delta\phi_1 + \cdots + \frac{\partial b_2^0}{\partial \phi_{253}}\Delta\phi_{253},$$

$$\widehat{b_2} = b_2^0 + (b_2^{\phi_1} - b_2^0 + \cdots + (b_2^{\phi_{253}} - b_2^0)$$

$$= \sum_{i=1}^{253} b_2^{\phi_i} - (252)b_2^0.$$

Expanding on step c, let $h^{\phi_i}(\bullet) = (h_1^0(\bullet), \ldots, h_i^{\phi_i}(\bullet), \ldots, h_{253}^0(\bullet))$ and $\boldsymbol{\varphi}^{\phi_i} = (\varphi_1^0, \ldots, \varphi_i^{\phi_i}, \ldots, \varphi_{253}^0)$, where $h_i^{\phi_i}(\bullet)$ denotes the spline fit for player $i$, adjusted for random effects $b_2$ and $b_3$, while allowing the first-order autocorrelation parameter $\phi_i$ to be determined freely (with all other values of $\phi$ set to 0) and $\varphi_i^{\phi_i}$ denotes the corresponding vector of autocorrelated error components for player $i$. Then $b_2^{\phi_i}$ and $b_3^{\phi_i}$ are the solution vectors from estimating $\mathbf{s}_2 = \mathbf{R}b_2^{\phi_i} + \mathbf{C}^i b_3^{\phi_i} + \boldsymbol{\varepsilon}$, where $\mathbf{s}_2 = \mathbf{s} - \mathbf{P}(h^{\phi_i}(\bullet) + \boldsymbol{\varphi}^{\phi_i})$ and $\mathbf{C}^i$ is an $N \times (m_i + 1)$ matrix of $m_i$ player–course interactions for player $i$ and one additional interaction for all other players. By defining $\mathbf{C}^i$ in this way, rather than in terms of the full $N \times 12{,}485$ matrix of 253 groups of nested player–course interactions, the time required for step c is reduced from two days to three hours. Finally, let $b_3^i$ denote the first $m_i$ rows of the solution vector $b_3^{\phi_i}$. With these definitions, we estimate the vector of starting values of random player–course effects as $\widehat{b_3} = b_3^0 + (b_3^1, \ldots, b_3^{253})'$.

All cubic splines are estimated in R using maximum likelihood in connection with the ssr function in the assist package of Wang and Ke (2004). Random round–course and player–course effects are estimated in R using REML as implemented in the lmer function of Pinherio and Bates contained in the Matrix package. (We obtain essentially identical results using a maximum likelihood specification in the lmer function.) After obtaining initial estimates of $b_2$ and $b_3$ in step 1, we employ seven iterations of steps 2–6. At the seventh iteration, the maximum absolute change from the sixth iteration in the solution values for the 848 round–course effects and 12,849 player–course effects are .00117 and .00019, respectively. Also, the maximum absolute change in the 64,364 point estimates over all 253 splines is .00199. All of these summary values decrease monotonically in steps 1–7.

*[Received October 2003. Revised June 2006.]*

## REFERENCES


Albright, S. C. (1993), "A Statistical Analysis of Hitting Streaks in Baseball," *Journal of the American Statistical Association*, 88, 1175–1183.

Berry, S. M. (2001), "How Ferocious Is Tiger?" *Chance*, 14, 51–56.

Burns, P. (2002), "Robustness of the Ljung–Box Test and Its Rank Equivalent," working paper, available at *http://www.burns-stat.com*.

Clark, R. D., III (2002a), "Do Professional Golfers 'Choke'?" *Perceptual and Motor Skills*, 94, 1124–1130.

——— (2002b), "Evaluating the Phenomenon of Choking in Professional Golfers," *Perceptual and Motor Skills*, 95, 1287–1294.

——— (2003a), "Streakiness Among Professional Golfers: Fact or Fiction?" *International Journal of Sports Psychology*, 34, 63–79.

——— (2003b), "An Analysis of Streaky Performance on the LPGA Tour," *Perceptual and Motor Skills*, 97, 365–370.

——— (2004a), "On the Independence of Golf Scores for Professional Golfers," *Perceptual and Motor Skills*, 98, 675–681.

——— (2004b), "Do Professional Golfers Streak? A Hole-to-Hole Analysis," *in Proceeding of the Statistics in Sports Section*, American Statistical Association, pp. 3207–3214.

Efron, B., and Tibshirani, R. J. (1998), *An Introduction to the Bootstrap*, Boca Raton, FL: Chapman & Hall/CRC.

Gilden, D., and Wilson, S. (1995), "Streaks in Skilled Performance," *Psychonomic Bulletin & Review*, 2, 260–265.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall.

Klaassen, F., and Magnus, J. (2001), "Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model," *Journal of the American Statistical Association*, 96, 500–509.

Ljung, G. M. (1986), "Diagnostic Testing of Univariate Time Series Models," *Biometrika*, 73, 725–730.

Ljung, G. M., and Box, G. E. P. (1978), "On a Measure of Lack of Fit in Time Series Models, *Biometrika*, 65, 297–303.

Mendenhall, W., Scheafler, R. L., and Wackerly, D. D. (1981), *Mathematical Statistics With Applications*, Boston: Duxbury Press.

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479–498.

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the $q$-Value," *The Annals of Statistics*, 31, 2013–2035.

Wang, Y. (1998), "Smoothing Spline Models With Correlated Random Errors," *Journal of the American Statistical Association*, 93, 341–348.

Wang, Y., and Ke, C. (2004), *ASSIST: A Suite of S Functions Implementing Spline Smoothing Techniques*, [online] manual. Available at *http://www.pstat.ucsb.edu/faculty/yuedong/ASSIST/manual/assist.html*.

Wang, Y., and Wahba, G. (1995), "Bootstrap Confidence Intervals for Smoothing Spline Estimates and Their Comparison to Bayesian Confidence Intervals," *Journal of Statistical Computation and Simulation*, 51, 263–279.